

# LD-DETR: 用于视频时刻检索和精彩片段检测的循环解码器检测 Transformer

赵鹏程, 贺之纤, 张富为, 林淑金\*, 周凡

中山大学, 中国广州

qingchen239@gmail.com, hezhx29@mail2.sysu.edu.cn,

zhangfw5@mail2.sysu.edu.cn, linshjin@mail.sysu.edu.cn, isszf@mail.sysu.edu.cn

## 摘要

视频时刻检索与精彩片段检测旨在根据文本查询找到视频中的对应内容。现有模型通常首先使用对比学习方法来对齐视频和文本特征, 然后融合并提取多模态信息, 最后使用 Transformer 解码器解码多模态信息。然而, 现有方法面临几个问题: (1) 数据集中不同样本之间重叠的语义信息阻碍了模型的多模态对齐性能; (2) 现有模型无法有效提取视频的局部特征; (3) 现有模型使用的 Transformer 解码器无法充分解码多模态特征。针对上述问题, 我们提出了用于视频时刻检索和精彩片段检测任务的 LD-DETR 模型。具体而言, 我们首先将相似度矩阵提取到单位矩阵以减轻重叠语义信息的影响。然后, 我们设计了一种方法, 使卷积层能够更有效地提取多模态局部特征。最后, 我们将 Transformer 解码器的输出反馈到其自身中, 以充分解码多模态信息。我们在四个公共数据集上对 LD-DETR 进行了评估, 并进行了广泛的实验, 以证明我们的方法的优越性和有效性。我们的模型在 QVHighlight、Charades-STA 和 TACoS 数据集上的表现优于最先进的模型 (State-Of-The-Art)。我们的代码可在以下网址获取: <https://github.com/qingchen239/ld-detr>

## 1 介绍

视频时刻检索旨在识别视频中与给定文本查询相对应的特定时刻 (Liu et al. 2015; Anne Hendricks et al. 2017; Gao et al. 2017; Liu et al. 2018; Escorcia et al. 2019b)。精彩片段检测评估不同时间片段与文本的相关程度 (Yao, Mei, and Rui 2016; Zhang et al. 2016; Gygli, Song, and Cao 2016; Yu et al. 2018; Xiong et al. 2019)。随着数字设备和平台的发展, 用户对于视频内容的需求大幅增加, 快速准确地找到视频中有趣的片段成为重要需求, 因此视频时刻检索与精彩片段检测的研究受到广泛关注。

现有的视频时刻检索和精彩片段检测模型往往使用对比学习来对齐视频和文本特征 (Moon et al. 2023b;

查询: 两个十几岁的男孩在机场大厅里。Two teen boys are in a airport lobby.

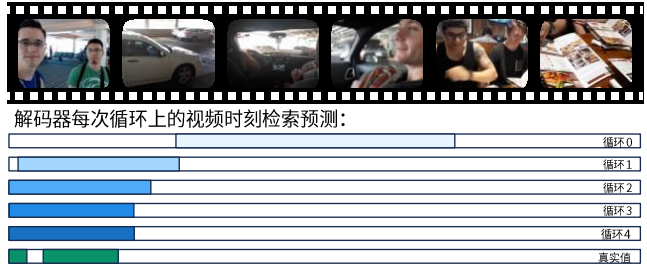


图 1: 循环解码器使视频时刻检索更加准确。我们将每次循环时循环解码器输出对应的视频时刻检索结果可视化。随着循环次数的增加, 结果越来越接近真实值。此实验以 M-DETR (Lei, Berg, and Bansal 2021) 为基准模型。

Sun et al. 2024; Moon et al. 2023a; Liu et al. 2024b; Xiao et al. 2024), 使用注意力机制来融合和提取多模态信息 (Liu et al. 2022b; Moon et al. 2023b; Sun et al. 2024), 并使用 Transformer 解码器和一个作为查询 (query) 的零矩阵来解码融合的多模态信息 (Zheng et al. 2020; Lei, Berg, and Bansal 2021)。

然而, 现有方法面临几个问题: (1) 在对比学习中, 方法通常将来自不同样本的特征视为负样本 (Sun et al. 2024; Moon et al. 2023a), 但相同的语义信息不可避免地出现在不同的样本中 (例如, “人吃饭” 和 “人喝水” 都具有相同的信息 “人”) (Jung et al. 2023)。将它们简单地视为完全的负样本会阻碍多模态对齐的性能。(2) 相应的内容通常是视频的一小部分, 具有很强的局部相关性, 但当前的模型忽略了提取视频的局部特征。一种直观的方法是使用卷积层 (LeCun et al. 1998; Krizhevsky, Sutskever, and Hinton 2012) 来提取局部特征 (Xiao et al. 2024; 黎金宇 2024), 但简单地堆叠卷积层并不能提高现有模型的性能。(3) 研究表明, Transformer 解码器不足以处理检索 (Yang et al. 2024; Liu et al. 2022a; Zhu et al. 2020; Gao et al. 2021b; Meng et al. 2021; Wang et al. 2022a; Yao et al. 2021)。

\*通讯作者。

更大的解码器可能会提高模型的检索能力，但也存在过拟合的风险。

为了解决这些问题，我们提出了循环解码器检测 Transformer 模型 (Loop Decoder DETection TRansformer, LD-DETR): (1) 与其他以单位矩阵作为目标的对比学习方法不同，我们将一个表示样本间相关性的矩阵提取到单位矩阵中，以减轻语义信息重叠的影响。(2) 卷积层的感受野较小，这一特性使得网络能够捕获局部信息。我们设计了一种方法，使得堆叠的卷积层能够更高效地提取多模态局部特征。(3) 研究表明，当 Transformer 解码器的查询 (*query*) 携带目标信息时，解码器能够更好地解码输入信息 (Liu et al. 2022a)，而 Transformer 解码器的输出也恰恰携带目标信息。受此启发，我们将 Transformer 解码器的输出作为查询反馈到自身，如图 1 所示，以增强其充分解码多模态融合信息的能力，同时又不增加过拟合的风险。

我们在四个公共数据集上对 LD-DETR 进行了评估，并进行了大量实验以证明我们方法的优越性和有效性。

我们工作的主要贡献总结如下：

- 我们提出了一种即插即用的方法提取对齐 (Distill Align)，该方法在对齐多模态特征时考虑训练样本之间重叠语义信息的影响，以提高模型的性能。
- 我们引入了卷积融合器 (Convolutional Fuser) 来更好地提取视频中的局部特征并取得了优异的效果。
- 我们提出了一种即插即用的方法，循环解码器 (Loop Decoder)，它提高了解码器充分解码多模态融合信息的能力，而不会导致过拟合。
- 基于以上方法，我们设计了用于视频时刻检索和精彩片段检测任务的模型 LD-DETR，并在多个数据集上验证了其先进性和有效性。我们的模型在 QVHighlight、Charades-STA 和 TACoS 数据集上的表现优于最先进的模型。

## 2 相关工作

自 QVHighlight 数据集 (Lei, Berg, and Bansal 2021) 提出以来，视频时刻检索和亮点检测任务被联合研究，并提出了许多基于检测 Transformer (DEtection TRansformer, DETR) (Zheng et al. 2020) 的模型。这些方法主要从对齐多模态特征、融合提取多模态特征、解码多模态信息三个角度对模型进行改进。

### 2.1 对比学习和对齐多模态特征

对比学习是一种通过比较和对比不同样本来提高其区分特征和识别模式的性能的机器学习方法。CMC (Tian, Krishnan, and Isola 2020) 利用对比学习将同一幅图像的不同视角映射到相似的语义空间中，证明了对齐多模态信息的可行性。MoCo (He et al. 2020)

通过动量更新编码器获取负样本的特征并将其存储在队列中，以增加参与对比学习的样本数量。CLIP (Radford et al. 2021) 为每个图像和文本提取全局特征，通过确保它们的相关矩阵近似于单位矩阵来对齐它们。

现有的视频时刻检索和精彩片段检测模型一般采用类似 CLIP 的方法对视频和文本特征进行对齐。TR-DETR (Sun et al. 2024) 将单模态编码器编码后的视频和文本特征的平均值作为样本的全局特征，再采用类似 CLIP 的方法对多模态特征进行对齐，最后再进行混合。CG-DETR (Moon et al. 2023a) 采用两个不同的编码器分别提取正样本和负样本的视频和文本的全局特征，然后通过类似 CLIP 的方法分别进行对齐。在这些类似 CLIP 的方法中，对比学习所涉及的特征数量受到批次大小 (batch size) 的限制。

BM-DETR (Jung et al. 2023) 提出了弱对齐问题，数据集中不同样本之间重叠的语义信息降低了模型性能，并在多模态特征融合层面解决该问题。我们认为该问题也可能降低多模态对齐，我们将在多模态对齐层面解决该问题。

### 2.2 融合和提取多模态特征

Moment-DETR (Lei, Berg, and Bansal 2021) 只是将视频和文本特征拼接起来，然后送入 Transformer 编码器。UMT (Liu et al. 2022b) 和 QD-DETR (Moon et al. 2023b) 提出了交叉注意力 Transformer 编码器，利用文本特征对视频特征进行编码，以去除视频特征中与查询文本无关的信息。CG-DETR (Moon et al. 2023a) 更进一步，在交叉注意力 Transformer 编码器的基础上将噪声拼接到文本特征上，以更好地去除无关的信息。TR-DETR (Sun et al. 2024) 引入了视觉特征细化来过滤掉无关的视频信息。现有的这些方法只注重去除视频特征中与文本无关的特征，而忽略了视频本身的时间结构，将提取视频特征的任务留给了 Transformer 编码器 (Vaswani et al. 2017)。但 Transformer 编码器的全局注意力计算 (Bahdanau, Cho, and Bengio 2014) 稀释了局部细节的权重，导致模型忽略了视频局部特征的提取。UVCOM (Xiao et al. 2024) 提出了综合集成模块方法，其中有一个卷积层来提取局部多模态特征。CDIM (黎金字 2024) 还提出了一种跨模态卷积交互方法，该方法堆叠扩张卷积层以增强模型的感知能力。

### 2.3 解码多模态信息

现有的模型大多基于一个应用于物体检测的模型 DETR (Zheng et al. 2020)，研究 (Yang et al. 2024; Liu et al. 2022a; Zhu et al. 2020; Gao et al. 2021b; Meng et al. 2021; Wang et al. 2022a; Yao et al. 2021) 表明 Transformer 解码器对融合多模态信息的处理不够充分，并提出了适合物体检测任务的解决方案。MomentDiff (Li et al. 2024) 注意到训练样本分布不均导

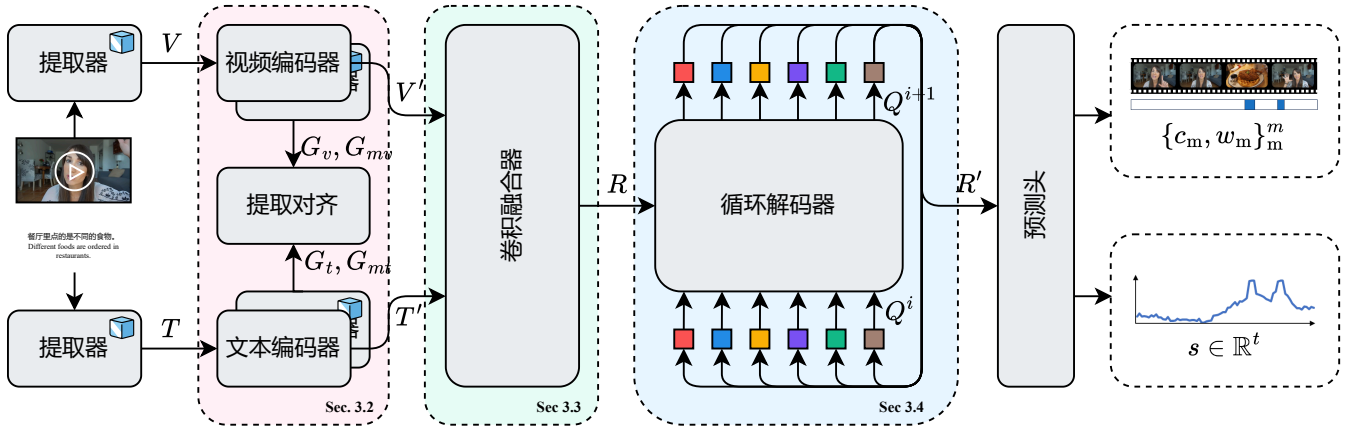


图 2: 我们的模型 LD-DETR 的总体框架。对于标有冰块的方法，其参数在训练过程中不受梯度下降的影响。

致模型泛化能力不足，在解码器中引入扩散模型的方法解决这个问题。UVCOM (Xiao et al. 2024) 利用文本和视频特征的交叉注意作为 Transformer 解码器的查询。TaskWeave (Yang et al. 2024) 使用类似 DAB-DETR (Liu et al. 2022a) 的解码器解码视频时刻检索特征，使用另一个解码器解码精彩片段检测特征。

### 3 模型和方法

#### 3.1 LD-DETR 概述

给定一个包含  $t$  个片段 (clip) 的视频和一个包含  $n$  个词 (token) 的文本查询，视频时刻检索和精彩片段检测的目标是找到与视频中的文本查询相关的所有时刻  $\{c_m, w_m\}_m^m$ ，其中  $c_m$  和  $w_m$  表示第  $m$  个时刻的中心时间和持续时间长度， $m$  是预测时刻的总数，并预测片段级别的所有时刻的显著性得分  $s \in \mathbb{R}^t$ 。

如图 2 所示，LD-DETR 可分为五个部分：单模态编码器 (Unimodal Encoder)、提取对齐 (Distill Align)、卷积融合器 (Convolutional Fuser)、循环解码器 (Loop Decoder) 和预测头 (Prediction Heads)。

输入的视频和文本首先被输入到预先训练的特征提取器中，以提取视频和文本特征  $V \in \mathbb{R}^{b \times n \times d_v}$ ,  $T \in \mathbb{R}^{b \times n \times d_t}$ ，其中  $b$  是批量大小， $d_v$  和  $d_t$  分别是提取的视频片段和文本标记特征的维度。然后将视频和文本特征输入到两个单模态编码器  $UE_v(\cdot), UE_t(\cdot)$  中，以映射到潜在空间  $V' \in \mathbb{R}^{b \times l \times d}$ ,  $T' \in \mathbb{R}^{b \times n \times d}$ ，其中  $d$  是模型的隐藏维度。我们使用两个单模态编码器和两个动量单模态编码器  $UE_{vm}(\cdot), UE_{tm}(\cdot)$  获取全局特征。将单模态编码器获取的两个全局特征和两个动量全局特征  $G_v, G_{mv}, G_t, G_{mt} \in \mathbb{R}^{b \times d}$  输入到提取对齐方法中，保证特征映射到同一个空间。然后将映射后的特征  $V', T'$  输入到卷积融合器中，得到多模态特征  $R \in \mathbb{R}^{b \times q \times d}$ 。然后将多模态特征  $R$  和零矩阵  $O \in \mathbb{R}^{b \times q \times d}$  一起送入循环解码器，得到解码特征  $R' \in \mathbb{R}^{b \times q \times d}$ ，其中  $q$  为超参数，表

示参考点的数量。最后将解码特征送入预测头，得到预测时间  $\{c_m, w_m\}_m^m$  和显著性得分  $s \in \mathbb{R}^t$ 。

#### 3.2 单模态编码器 (Unimodal Encoder) 和提取对齐 (Distill Align)

我们使用一个两层的多层感知器 (Rumelhart, Hinton, and Williams 1986) 作为单模态编码器  $UE(\cdot)$  将提取的特征  $X \in \mathbb{R}^{b \times x \times d_x}$  映射到潜在空间  $X' \in \mathbb{R}^{b \times x \times d}$ ，其中  $X \in \{V, T\}$  同时  $x \in \{t, n\}$ 。我们利用样本在潜在空间  $X'$  中在片段 (clip) 或词 (token) 维度上的特征的平均值来获得每个样本的全局特征  $G \in \mathbb{R}^{b \times d}$ 。

我们利用两个可学习的单模态编码器  $UE(\cdot) \in \{UE_v(\cdot), UE_t(\cdot)\}$  和两个动量单模态编码器  $UE_m(\cdot) \in \{UE_{mv}(\cdot), UE_{mt}(\cdot)\}$  将提取的特征  $V \in \mathbb{R}^{b \times x \times d_v}$ ,  $T \in \mathbb{R}^{b \times x \times d_t}$  映射到潜在空间，分别得到四个全局特征  $G_v, G_{mv}, G_t, G_{mt}$ 。动量单模态编码器由相应的单模态编码器更新：

$$UE_{m\theta}^0 = UE_{\theta}^0, \quad (1)$$

$$UE_{m\theta}^i = mUE_{m\theta}^{i-1} + (1-m)UE_{\theta}^i, \quad \text{when } i > 0, \quad (2)$$

其中  $m \in [0, 1]$  为动量 (momentum) 系数， $X_{\theta}$  表示方法  $X$  中的所有参数。得到映射特征  $V', V'_m, T', T'_m \in \mathbb{R}^{b \times x \times d}$  后，我们计算每个样本的全局特征  $G_v, G_{mv}, G_t, G_{mt} \in \mathbb{R}^{b \times d}$ 。我们将动量全局特征推送到动量全局特征队列  $Q_v, Q_t \in \mathbb{R}^{l \times d}$ ，其中  $l$  表示队列长度。

图 3 展示了提取对齐的结构。此时，我们有视频全局特征  $G_v \in \mathbb{R}^{b \times d}$ 、视频和文本动量全局特征  $G_{mv}, G_{mt} \in \mathbb{R}^{b \times d}$ ，以及文本动量全局特征队列  $Q_t \in \mathbb{R}^{l \times d}$ 。通过计算余弦相似度  $s(\cdot, \cdot)$ ，我们得到了这些全局特征中的视频到文本相似度矩阵  $S_{v2t}, S_{v2tm} \in \mathbb{R}^{b \times l}$ ：

$$S_{v2t} = s(G_v, Q_t), \quad (3)$$

$$S_{v2tm} = s(G_{mv}, Q_t). \quad (4)$$



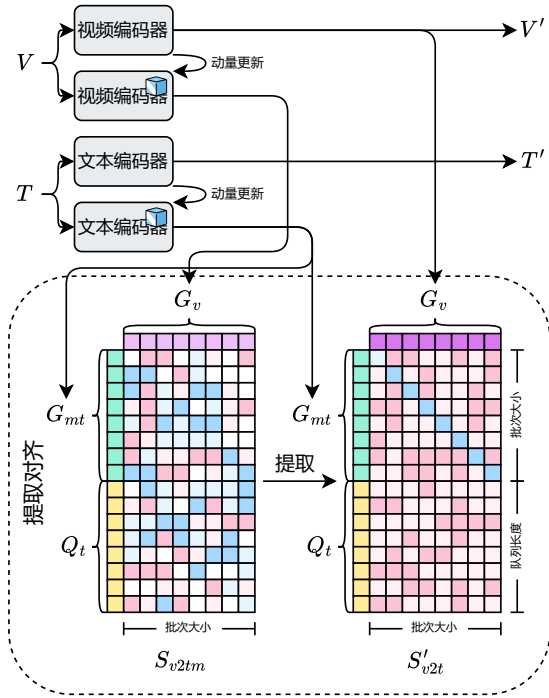


图 3: 提取对齐的结构。

为了减轻语义信息重叠的影响，我们将相似度矩阵  $S_{v2tm} \in \mathbb{R}^{b \times l}$  提取到单位矩阵  $I$ ，作为多模态比对的目标  $S'_{v2t} \in \mathbb{R}^{b \times l}$ ：

$$S'_{v2t} = \alpha S_{v2tm} + (1 - \alpha)I, \quad (5)$$

其中  $\alpha \in [0, 1]$  是提取系数 (distillation coefficient)。最后，视频到文本对齐损失  $\mathcal{L}_{v2t}$  是

$$\mathcal{L}_{v2t} = \text{CE}(S_{v2t}, S'_{v2t}), \quad (6)$$

其中  $\text{CE}(\cdot, \cdot)$  是交叉熵损失 (cross entropy loss)。

类似地，我们计算文本到视频的对齐损失  $\mathcal{L}_{t2v}$ 。最终的多模态对齐损失为

$$\mathcal{L}_{\text{align}} = (\mathcal{L}_{v2t} + \mathcal{L}_{t2v})/2. \quad (7)$$

### 3.3 卷积融合器 (Convolutional Fuser)

图 4 显示了卷积融合器的结构。首先，我们将映射到潜在空间的视频特征  $V' \in \mathbb{R}^{b \times t \times d}$  和文本特征  $T' \in \mathbb{R}^{b \times n \times d}$  输入到 V2T 提取器 (V2T Extractor) (Sun et al. 2024)、T2V 编码器 (T2V Encoder) (Moon et al. 2023b) 和 Transformer 编码器 (Vaswani et al. 2017)，以获得与文本无关的视频特征  $V'' \in \mathbb{R}^{b \times t \times d}$ 。随后，将与文本无关的视频特征  $V'' \in \mathbb{R}^{b \times t \times d}$  输入到卷积块 (Convolutional Blocks) 中。经过这个残差网络后，我们得到局部增强的多模态融合特征  $V''' \in \mathbb{R}^{b \times t \times d}$ 。最后，将局部增强的多模态融合特征  $V''' \in \mathbb{R}^{b \times t \times d}$  输入到另一个 Transformer 编码器 (Vaswani et al. 2017) 中，得到多模态特征  $R \in \mathbb{R}^{b \times t \times d}$ 。

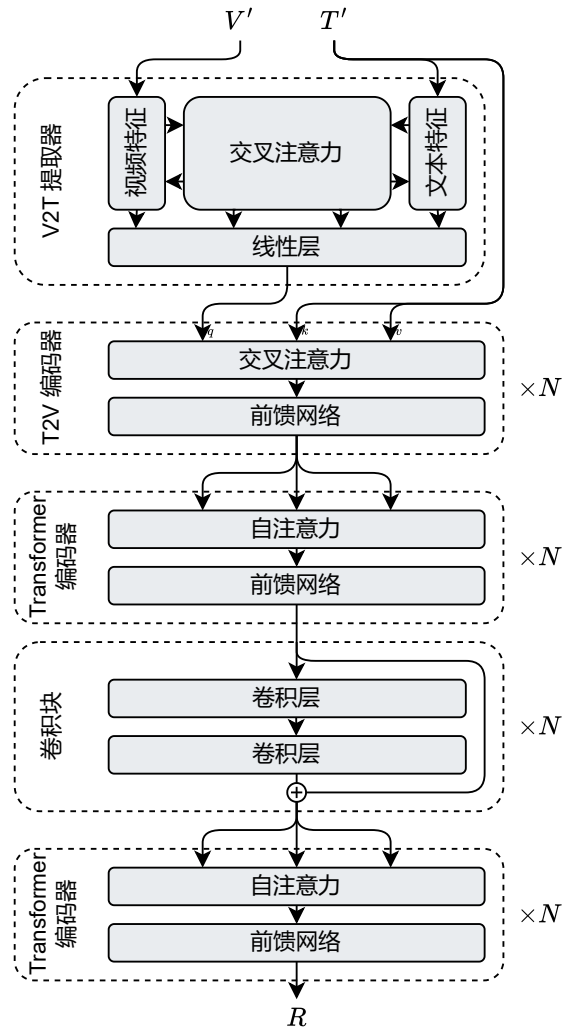


图 4: 卷积融合器的结构。

**V2T 提取器 (V2T Extractor)** 首先，我们利用映射到潜在空间的视频特征  $V' \in \mathbb{R}^{b \times t \times d}$  和文本特征  $T' \in \mathbb{R}^{b \times n \times d}$  得到视频特征与文本特征之间的相关矩阵  $A \in \mathbb{R}^{b \times t \times n}$ ：

$$A_1 = \text{Linear}(V'), \quad (8)$$

$$A_2 = \text{Linear}(T'), \quad (9)$$

$$A_3 = \text{Linear}(V')T'^T, \quad (10)$$

$$A = A_1 + A_2^T + A_3, \quad (11)$$

当矩阵相加时，较短的维度会自行扩展为与其他矩阵相同的长度。然后我们在文本和视频维度上对相关矩阵  $A$  执行 softmax，得到另外两个相关矩阵  $A_r, A_c \in \mathbb{R}^{b \times t \times n}$ 。

然后我们得到与文本无关的视频特征  $V'_v \in \mathbb{R}^{b \times t \times d}$ :

$$T_v = A_r T', \quad (12)$$

$$V_t = A_t A_c^T V', \quad (13)$$

$$V_{cat} = [V' || T_v || V' \circ T_v || V' \circ V_t], \quad (14)$$

$$V'_{cat} = \text{Linear}(V_{cat}), \quad (15)$$

$$B = \text{Softmax}(\text{Linear}(T')), \quad (16)$$

$$T_p = T'^T B, \quad (17)$$

$$V''_{cat} = [V'_{cat} || T_p], \quad (18)$$

$$V'_v = \text{Linear}(V''_{cat}), \quad (19)$$

其中  $[\cdot || \cdot]$  表示连接,  $\circ$  表示哈达玛积。

**V2T 编码器 (T2V Encoder)** 然后, 我们利用与文本无关的视频特征  $V'_v$  来获得文本引导的与文本无关的视频特征  $V'' \in \mathbb{R}^{b \times t \times d}$ :

$$Q_v = \text{Linear}(V'_v), \quad (20)$$

$$K_t = \text{Linear}(T'), \quad (21)$$

$$V_t = \text{Linear}(T'), \quad (22)$$

$$\text{Attention}(Q_v, K_t, V_t) = \text{Softmax}\left(\frac{Q_v K_t^T}{d}\right) V_t, \quad (23)$$

$$V'' = \text{FFN}(\text{Attention}(Q_v, K_t, V_t)). \quad (24)$$

**Transformer 编码器** 这里的 Transformer 编码器和其 他论文里的没什么区别 (Vaswani et al. 2017)。需要注意的是, 这里的两个 Transformer 编码器并不共享参数。

**卷积块 (Convolutional Blocks)** 随后, 将与文本无关的视频特征  $V'' \in \mathbb{R}^{b \times t \times d}$  输入卷积块。根据之前在图像识别中的工作, 我们使用类似于 ResNet (He et al. 2016) 的残差块  $\text{RB}(\cdot)$ :

$$X_i = \text{RB}_i(X_{i-1}) = \sigma(X_{i-1} + \mathcal{F}(X_{i-1})), \quad (25)$$

$$\mathcal{F}(X_{i-1}) = \text{BN}(\text{Conv}(\sigma(\text{BN}(\text{Conv}(X_{i-1}))))), \quad (26)$$

其中  $\text{Conv}(\cdot)$  表示卷积层,  $\text{BN}(\cdot)$  表示批量归一化,  $\sigma(\cdot)$  表示激活函数。我们堆叠  $N$  个残差块来提取视频特征中的局部信息:

$$X_0 = V'', \quad (27)$$

$$X_i = \text{RB}_i(X_{i-1}), \quad \text{when } i > 0, \quad (25)$$

$$V''' = X_N. \quad (28)$$

经过这个残差网络后, 我们得到局部增强的多模态融合特征  $V''' \in \mathbb{R}^{b \times t \times d}$ 。

### 3.4 循环解码器 (Loop Decoder)

研究表明, 当 Transformer 解码器的查询 (query) 携带目标信息时, 解码器能够更好地解码输入信息 (Liu

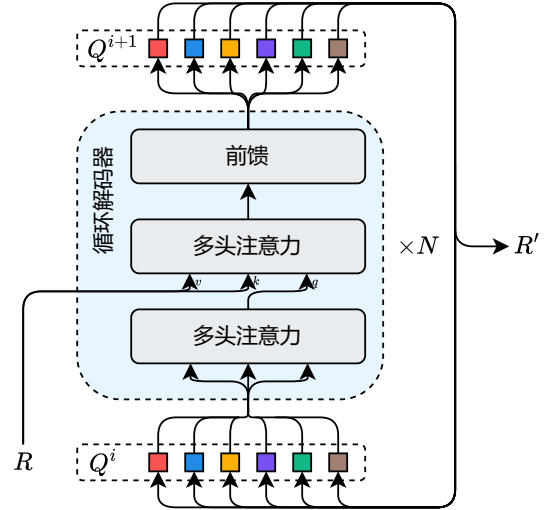


图 5: 循环解码器的结构。

et al. 2022a), 而 Transformer 解码器的输出也恰恰携带目标信息。受此启发, 我们将 Transformer 解码器的输出作为查询反馈到自身, 以增强其解码多模态融合信息的能力。

图 3 展示了循环解码器的结构。我们将一个零矩阵  $O \in \mathbb{R}^{b \times q \times d}$  和多模态特征  $R \in \mathbb{R}^{b \times t \times d}$  送入 Transformer 解码器  $\text{TD}(\cdot, \cdot)$  (Vaswani et al. 2017)。经过 Transformer 解码器的  $N$  次循环后, 得到解码特征  $R' \in \mathbb{R}^{b \times q \times d}$ :

$$Q^0 = O, \quad (29)$$

$$Q^i = \text{TD}(Q^{i-1}, R), \quad \text{when } i > 0, \quad (30)$$

$$R' = Q^N, \quad (31)$$

其中  $q$  为超参数, 代表参考点的数量。与更大的解码器相比, 循环解码器没有引入新的参数, 不会引起过拟合。

### 3.5 预测头 (Prediction Heads)

我们使用类似于 QD-DETR (Moon et al. 2023b) 和 TR-DETR (Sun et al. 2024) 的预测头。

**视频时刻检索预测头 (Video Moment Retrieval Prediction Head)** 我们将解码后的特征  $R' \in \mathbb{R}^{b \times q \times d}$  输入多层感知器, 得到  $q$  个预测时间  $\{c_m, w_m\}_m^q$ :

$$\{s, e\} = \text{MLP}(R'), \quad (32)$$

其中  $s$  和  $e$  分别表示一个预测时间的开始和结束。同时利用解码后的特征  $R$  和多模态特征  $R' \in \mathbb{R}^{b \times t \times d}$  得到每个预测时间  $p \in \mathbb{R}^q$  的置信度:

$$p = \text{MLP}(R') + \text{Sigmoid}^{-1}(\text{MLP}(R')), \quad (33)$$

其中  $\text{Sigmoid}^{-1}(\cdot)$  表示反 S 型函数。这样, 我们就得到了视频中与文本查询相关的所有时间  $\{c_m, w_m\}_m^m$ 。

**精彩片段检测预测头 (Highlight Detection Prediction Head)** 得到所有时间之后, 我们取出时间中所有片段 (clip) 对应的解码特征  $R'' \in \mathbb{R}^{b \times t' \times d}$ , 其中  $t'$  表示视频时刻检索预测中的片段总数, 将它们输入到门控循环单元 (Chung et al. 2014), 并使用隐藏状态作为视频  $G'_v \in \mathbb{R}^{b \times d}$  的新全局特征:

$$o = \text{GRU}(R''), \quad (34)$$

$$G'_v = \text{GRU}_\theta, \quad (35)$$

其中  $o$  表示我们不关心的门控循环单元的输出,  $\text{GRU}_\theta$  表示门控循环单元的隐藏状态。然后, 我们计算新的全局特征  $G'_v$  与每个剪辑的特征  $V'$  之间的相似度:

$$S = G'_v V'^T \quad (36)$$

最后得到显著性分数  $s \in \mathbb{R}^t$ :

$$M = \text{Linear}(R' \circ S + R'), \quad (37)$$

$$s = \text{sum}(M)/d, \quad (38)$$

其中  $\circ$  表示哈达玛积,  $\text{sum}(\cdot)$  表示对矩阵元素进行列求和,  $d$  是模型的隐藏维度。

### 3.6 目标损失

LD-DETR 的目标损失函数  $\mathcal{L}_{total}$  为

$$\mathcal{L}_{total} = \mathcal{L}_{mr} + \mathcal{L}_{hd} + \lambda_{align} \mathcal{L}_{align}, \quad (39)$$

其中

$$\mathcal{L}_{mr} = \mathcal{L}_{mom} + \lambda_{CE} \text{CE}(\hat{y}, y), \quad (40)$$

$$\mathcal{L}_{mom} = \lambda_{L1} ||m - \hat{m}|| + \lambda_{IoU} \mathcal{L}_{IoU}(m, \hat{m}), \quad (41)$$

$$\mathcal{L}_{hd} = \lambda_{marg} \mathcal{L}_{marg} + \lambda_{cont} \text{RAC}(X_r^{\text{pos}}, X_r^{\text{neg}}) \quad (42)$$

$$\mathcal{L}_{marg} = \max(0, \Delta + S(x^{\text{low}}) - S(x^{\text{high}})), \quad (43)$$

其中  $y$  和  $\hat{y}$  是前景或背景的真值及其对应的预测,  $m$  和  $\hat{m}$  是真值矩及其对应的预测,  $\text{IoU}$  损失  $\mathcal{L}_{IoU}(\cdot, \cdot)$  来自之前的研究 (Rezatofighi et al. 2019),  $\Delta$  是边距,  $S(\cdot)$  是显著性分数估计器,  $x^{\text{high}}$  和  $x^{\text{low}}$  分别是来自两对高秩和低秩剪辑的视频标记,  $\text{CE}(\cdot, \cdot)$  是交叉熵损失,  $\text{RAC}(\cdot, \cdot)$  是排名感知对比损失 (Hoffmann et al. 2022),  $X_r^{\text{pos}}$  和  $X_r^{\text{neg}}$  分别是比迭代索引排序高和低的样本集合。

## 4 实验

### 4.1 数据集

我们在四个流行的视频时刻检索和精彩片段检测公共数据集上评估了我们的模型: QVHighlights (Lei, Berg, and Bansal 2021) (Gao et al. 2017) (Regneri et al. 2013) 和 TVSum (Song et al. 2015)。

由于数据集限制了提交的次数, 在与模型的比较中, 我们进行了多次实验, 并给出了最有可能获得最佳结果的实验结果。

因为 QVHighlight (Lei, Berg, and Bansal 2021) 是目前唯一一个同时支持视频时刻检索和精彩片段检测的数据集。因此, 在消融实验中, 我们在 QVHighlight 数据集验证集上进行了所有的实验。我们每个实验进行了五次, 分别使用 1、23、456、7890 和 1,2345<sup>1</sup> 作为随机种子, 并给出了所有实验结果的平均值和方差。

### 4.2 指标

我们采用了与之前研究相同的评估指标。具体来说, 我们计算了  $\text{IoU}$  阈值  $\theta_{IoU} = 0.5$  和  $0.7$  的  $\text{Recall}@1$ ,  $\theta_{IoU} = 0.5$  和  $0.7$  的  $\text{mAP}$  (平均精度, mean average precision) 以及一系列阈值  $[0.5 : 0.05 : 0.95]$  的  $\text{mAP}$ , 用于 QVHighlights 上的视频时刻检索。精彩片段检测采用  $\text{mAP}$  和  $\text{HIT}@1$ , 其中正样本定义为显著性得分为  $v$ 。在 Charades-STA 和 TACoS 数据集上, 我们使用  $\text{Recall}@1$  和  $\theta_{IoU} = \{0.3, 0.5, 0.7\}$  和  $\text{mIoU}$  来衡量视频时刻检索性能。对于 TVSum, 分别采用  $\text{mAP}$  和  $\text{Top-5 mAP}$ 。

### 4.3 实验设置

在所有实验中, 我们采用 CLIP (Radford et al. 2021) 和 Slowfast (Feichtenhofer et al. 2019) 作为提取器来提取视频特征, 采用 CLIP (Radford et al. 2021) 提取文本特征。在部分实验中, 采用 PANN (Kong et al. 2020) 提取音频特征。

默认情况下, 我们的模型使用 AdamW 优化器 (Loshchilov 2017) 训练了 200 个 epoch, 学习率为  $1e-4$ , 批次大小为 32, 隐藏维度为 256, 队列长度为 6,5536, 动量 (momentum) 系数为 0.995, 提取系数 (distillation coefficient) 为 0.4, 卷积块 (Convolutional Blocks) 层数为 5, 参考点 (reference points) 数为 10, 循环解码器循环数为 3。在 Charade-STA 数据集中, 我们使用 4 层卷积块。在 TACoS 数据集中, 我们使用蒸馏系数为 0.7。在 Charade-STA 数据集中, 我们使用提取系数为 0.3 训练了 100 个 epoch。在 TVSum 数据集中, 我们尝试了每种方法并更改了每个超参数以获得更好的结果。由于此数据集的特性, 每次实验中的超参数都不同。因此, 我们没有记录它们。

### 4.4 与其他模型的比较

图 6 直观地展示了 LD-DETR 与其他模型的比较。表 1 报告了 LD-DETR 在 QVHighlight 数据集上联合视频时刻检索和精彩片段检测任务中的表现。表中给出的所有模型均未使用预训练。表中的模型根据使用的提取器分为三类。我们的模型优于所有现有模型, 甚至

<sup>1</sup>为方便阅读, 我们在本论文中使用 “,” 作为数字分位符, 将数字以四位为一段进行分隔。

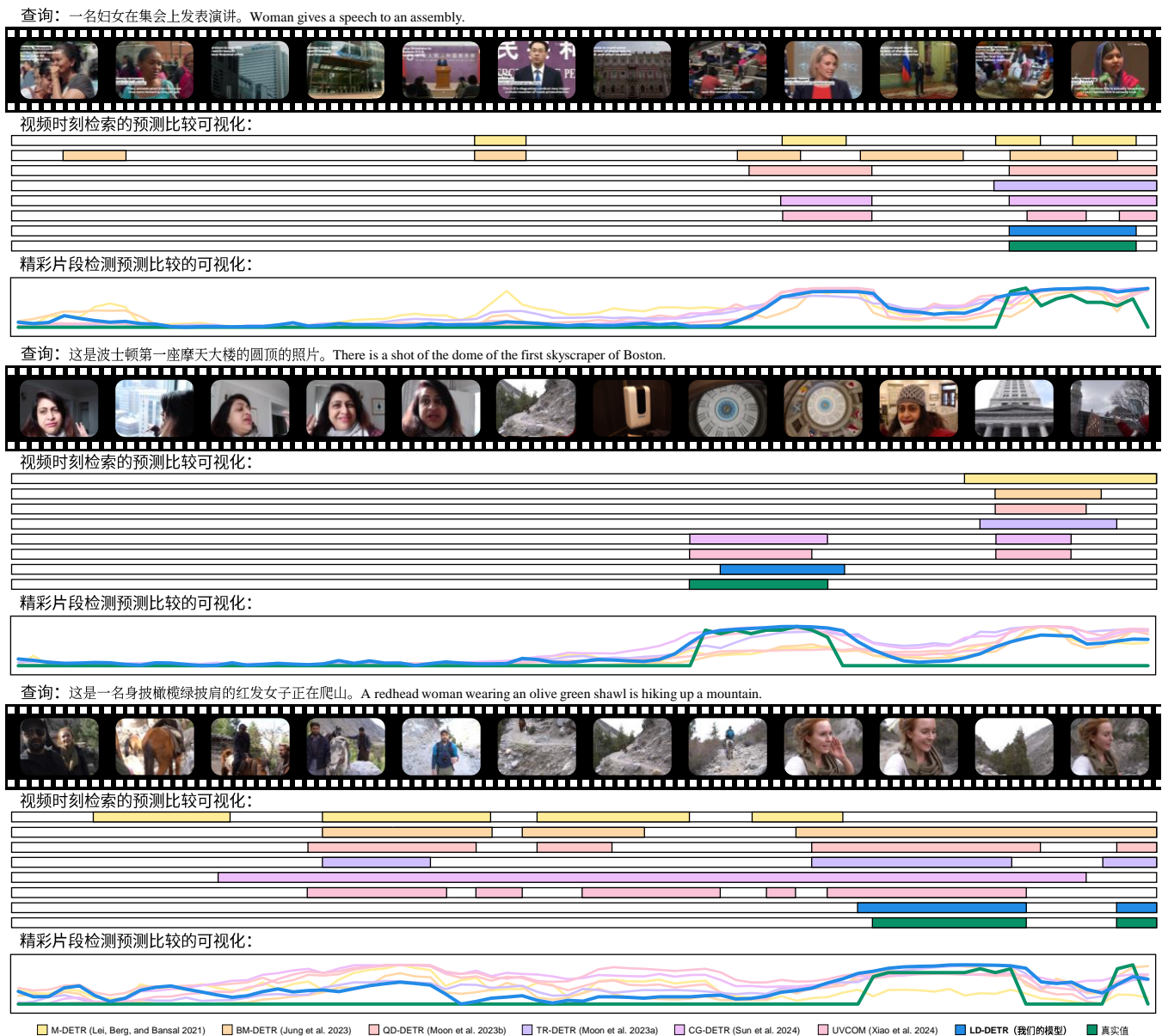


图 6: 视频时刻检索和精彩片段检测的可视化比较。

优于那些使用更多提取特征的模型。表 2 报告了 LD-DETR 在 TACoS 数据集和 CharadesSTA 数据集上视频时刻检索的表现。得益于我们提出的方法, LD-DETR 模型在 QVHighlight、Charades-STA 和 TACoS 数据集上的表现优于最先进的模型。

表 3 报告了 LD-DETR 在 TV-Sum 数据集上进行高光检测的表现。TV-Sum 数据集太小, 每个类别只有 4 个训练样本和 1 个测试样本。在训练过程中, 模型很快过拟合。即便如此, LD-DETR 模型在 TV-Sum 数据集上仍然取得了不错的效果。

#### 4.5 消融实验

**提取对齐的消融实验** 表 4 展示了提取对齐在多个模型上的表现。它展示了提取对齐作为一种即插即用的方法, 可以提高多个模型的性能。通过提取对齐方法, 将映射到潜在空间的视频和文本特征对齐到同一个语义空间, 并在对齐过程中考虑到不同训练样本中重叠的语义信息。

图 7 和表 5 显示, 提取对齐方法能够使模型取得较好的效果, 同时不会占用过多的 GPU 内存。可以看出, 在使用更大的批次大小时, 随着批次大小的增大, 模型的效果会越来越好。但是, 批次大小越大, 模型占用的



模型	视频时刻检索					精彩片段检测	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
<b>Slowfast + CLIP</b> (5.5 GB)							
BeautyThumb (Song et al. 2016)	-	-	-	-	-	14.36	22.88
DVSE (Liu et al. 2015)	-	-	-	-	-	18.75	21.79
MCN (Anne Hendricks et al. 2017)	11.41	2.72	24.94	8.22	10.67	-	-
CAL (Escorcia et al. 2019a)	25.49	11.54	23.40	7.65	9.89	-	-
XML (Lei et al. 2020)	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML <sup>+</sup> (Lei, Berg, and Bansal 2021)	46.69	33.46	47.89	34.67	34.90	35.38	55.06
Moment-DETR (Lei, Berg, and Bansal 2021)	52.89	33.02	54.82	29.40	30.73	35.69	55.60
Localizer (Yu et al. 2024)	54.50	36.50	-	-	32.30	-	-
UniVTG (Lin et al. 2023)	58.86	40.86	57.60	35.59	35.47	38.20	60.96
MomentDiff (Li et al. 2024)	57.42	39.66	54.02	35.73	35.95	-	-
VMRNet (缪翌, 张卫锋, and 徐领 2024)	59.94	42.84	55.56	37.75	36.87	-	-
LLaViLo (Ma et al. 2023)	59.23	41.42	9.72	-	36.94	-	-
MH-DETR (Xu et al. 2024)	60.05	42.48	60.75	38.13	38.38	38.22	60.51
QD-DETR (Moon et al. 2023b)	62.40	44.98	62.52	39.88	39.86	38.94	62.40
CDIM (黎金宇 2024)	60.51	45.53	61.36	41.05	39.94	37.69	60.05
BM-DETR (Jung et al. 2023)	60.12	43.05	63.08	40.18	40.08	-	-
MESM (Liu et al. 2024c)	62.78	45.20	62.64	41.45	40.68	-	-
EaTR (Jang et al. 2023)	61.36	45.79	61.86	41.91	41.74	37.15	58.65
LMR (Liu et al. 2024a)	64.40	47.21	64.65	43.16	42.56	-	-
TR-DETR (Sun et al. 2024)	64.66	48.96	63.98	43.73	42.62	39.91	63.42
CG-DETR (Moon et al. 2023a)	65.43	48.38	64.51	42.77	42.86	<u>40.33</u>	<b>66.21</b>
CDNet (Ma et al. 2024)	67.74	49.55	63.82	42.30	42.76	39.84	66.52
UVCOM (Xiao et al. 2024)	63.55	47.47	63.37	42.67	43.18	39.74	64.20
SFABD (Huang et al. 2024)	-	-	62.38	44.39	43.79	-	-
LLMEPET (Jiang et al. 2024)	66.73	49.94	65.76	43.91	44.05	40.33	65.69
UniVTG <sup>+</sup> (Chen et al. 2024)	66.65	<b>52.19</b>	64.37	<u>46.68</u>	45.18	40.18	64.77
BAM-DETR (Lee and Byun 2023)	62.71	48.64	64.57	46.33	<u>45.36</u>	-	-
TaskWeave (Yang et al. 2024)	64.26	50.06	65.39	46.47	45.38	39.28	63.68
<b>LD-DETR (我们的模型)</b>	<b>66.80</b>	<u>51.04</u>	<b>67.61</b>	<b>46.99</b>	<b>46.41</b>	<b>40.51</b>	65.11
<b>Slowfast + CLIP + PANN</b> (11.7 GB)							
UMT (Liu et al. 2022b)	56.23	41.18	53.38	37.01	36.12	38.18	59.99
VCSJT (Zhou et al. 2024)	59.14	42.02	55.76	37.79	36.37	38.59	62.45
MIM (Li et al. 2023)	59.99	41.50	55.85	36.84	36.45	38.96	62.39
LSJT (Wang et al. 2024)	60.51	41.50	56.33	36.70	36.66	39.13	61.22
MomentDiff (Li et al. 2024)	58.21	41.48	54.57	37.21	36.84	-	-
QD-DETR (Moon et al. 2023b)	63.06	45.10	63.04	40.10	40.19	39.04	62.87
UVCOM (Xiao et al. 2024)	<u>63.81</u>	<u>48.70</u>	<u>64.47</u>	<u>44.01</u>	<u>43.27</u>	<u>39.79</u>	<u>64.79</u>
<b>LD-DETR (我们的模型)</b>	<b>65.76</b>	<b>50.71</b>	<b>66.06</b>	<b>46.62</b>	<b>45.85</b>	<b>41.00</b>	<b>67.06</b>
<b>CLIP<sup>+</sup></b> (233.7 GB)							
R <sup>2</sup> -Tuning (Liu et al. 2024b)	68.03	49.35	69.04	47.56	46.17	40.75	64.20

表 1: QVHighlights 测试集上的联合视频时刻检索和高光检测结果。该表按使用的提取器对模型进行分类，并标明提取器和提取的特征大小。每列中测试集上每类特征中的最佳结果以**粗体**突出显示，第二佳结果以下划线突出显示。灰色显示的模型仅报告其在论文中的验证集上的结果。

GPU 内存也会越大，并且随着批次大小的增大，模型收敛的难度也会越大。具体来说，在批次大小 = 1,024 时，我们进行了 5 次实验，其中只有一次模型成功收

敛。而使用提取对齐时，随着队列长度的增加，模型的效果不断变好，而占用的 GPU 内存却小了很多。

图 8 和表 6 展示了提取对齐中提取系数对多模态



模型	TACoS				Charades-STA			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
CTRL (Gao et al. 2017)	18.32	13.30	-	-	-	23.63	8.89	-
ABLR (Yuan, Mei, and Zhu 2019)	19.50	9.40	-	13.40	-	-	-	-
SM-RL (Wang, Huang, and Wang 2019)	-	20.25	15.95	-	-	24.36	11.17	-
TGN (Chen et al. 2018)	21.77	18.90	-	-	-	-	-	-
SAP (Chen and Jiang 2019)	-	18.24	-	-	-	27.42	13.36	-
MIM† (Li et al. 2023)	-	-	-	-	-	43.92	25.89	-
MAN (Zhang et al. 2019)	-	-	-	-	-	46.53	22.72	-
FMAN (蒋寻 et al. 2023)	-	-	-	-	-	51.40	25.05	38.23
LSJT† (Wang et al. 2024)	-	-	-	-	-	44.62	25.13	-
DRN (Zeng et al. 2020)	-	23.17	-	-	-	45.40	26.40	-
UMT† (Liu et al. 2022b)	-	-	-	-	-	48.31	29.25	-
VCSJT† (Zhou et al. 2024)	-	-	-	-	-	51.21	30.22	-
SFABD (Huang et al. 2024)	-	-	-	-	-	-	30.51	-
BPNet (Xiao et al. 2021)	25.96	20.96	14.08	19.53	65.48	50.75	31.64	46.34
M-DETR (Lei, Berg, and Bansal 2021)	-	-	-	-	-	53.63	31.37	-
SCDM (Yuan et al. 2019b)	26.11	21.17	-	-	-	54.44	31.37	-
DCL (Nan et al. 2021)	38.84	29.07	19.05	28.26	67.63	50.24	32.88	48.02
HUAL (Ji et al. 2023)	-	-	-	-	70.40	52.69	28.90	48.11
VSLNet (Zhang et al. 2020a)	29.61	24.27	20.03	24.11	70.46	54.19	35.22	50.02
2D-TAN (Zhang et al. 2020b)	37.29	25.32	-	-	-	39.70	23.31	-
MMN (Wang et al. 2022b)	39.24	26.17	-	-	-	47.31	27.28	-
CrossGraphAlign (陈卓 et al. 2020)	39.80	26.40	-	-	-	-	-	-
CBLN (Liu et al. 2021)	38.98	27.65	-	-	-	61.13	38.22	-
CPNet (Li, Guo, and Wang 2021)	42.61	28.29	-	28.69	-	60.27	38.74	-
FVMR (Gao et al. 2021b)	41.48	29.12	-	-	-	59.46	35.48	-
SimVTP (Ma et al. 2022)	43.10	30.30	-	-	-	44.70	26.30	-
RaNet (Gao et al. 2021a)	43.34	33.54	-	-	-	60.40	39.65	-
MomentDiff (Li et al. 2024)	44.78	33.68	-	-	-	55.57	32.42	-
LLaViLo (Ma et al. 2023)	-	-	-	-	-	55.72	33.43	-
TaskWeave (Yang et al. 2024)	-	-	-	-	-	56.51	33.66	-
QD-DETR† (Moon et al. 2023b)	-	-	-	-	-	55.51	34.17	-
LMR (Liu et al. 2024a)	-	-	-	-	-	55.91	35.19	-
VLG-Net (Soldan et al. 2021)	45.46	34.19	-	-	-	-	-	-
GVL (Wang et al. 2023)	45.92	34.57	-	32.48	-	-	-	-
TR-DETR (Sun et al. 2024)	-	-	-	-	-	57.61	33.52	-
UniVTG (Lin et al. 2023)	51.44	34.97	17.35	33.60	70.81	58.01	35.65	50.10
UniVTG <sup>+</sup> (Chen et al. 2024)	-	-	-	-	68.06	57.18	36.05	-
BM-DETR (Jung et al. 2023)	50.31	35.42	-	-	-	59.48	38.33	-
SFEN (杨金福 et al. 2022)	47.30	36.10	-	-	-	-	-	-
MATN (Zhang et al. 2021)	48.79	37.57	-	-	-	-	-	-
这个模型没有名字 (Panta et al. 2024)	49.77	37.99	-	-	-	-	-	-
CDNet (Ma et al. 2024)	54.11	35.35	20.34	33.76	71.25	58.09	36.53	-
MS-DETR (Jing et al. 2023)	47.66	37.36	25.81	35.09	68.68	57.72	37.40	50.12
CG-DETR (Moon et al. 2023a)	52.23	39.61	22.23	36.48	70.43	58.44	36.34	50.13
UVCOM (Xiao et al. 2024)	-	36.39	23.32	-	-	59.25	36.64	-
R <sup>2</sup> -Tuning (Liu et al. 2024b)	49.71	38.72	25.12	35.92	70.91	59.78	37.02	50.86
LLMEPET (Jiang et al. 2024)	52.73	-	22.78	36.55	70.91	-	36.49	50.25
UnLoc (Yan et al. 2023)	-	-	-	-	-	60.80	38.40	-
MESM (Liu et al. 2024c)	52.69	39.52	-	36.94	-	61.24	38.04	-
MCMN (Han et al. 2023)	50.24	36.78	-	-	-	<b>62.69</b>	<u>41.38</u>	-
BAM-DETR (Lee and Byun 2023)	<u>56.69</u>	<u>41.54</u>	<b>26.77</b>	<u>39.31</u>	<u>72.93</u>	59.95	39.38	<u>52.33</u>
<b>LD-DETR (我们的模型)</b>	<b>57.61</b>	<b>44.31</b>	<u>26.24</u>	<b>40.30</b>	<b>73.06</b>	<u>62.28</u>	<b>42.23</b>	<b>53.14</b>

表 2: TACoS 和 Charades-STA 上的视频时刻检索结果。每列中的最佳结果以**粗体**突出显示，第二佳结果以下划线突出显示。† 表示使用音频模态进行训练。

模型	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
sLSTM (Zhang et al. 2016)	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
SG (Yuan et al. 2019a)	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6	46.2
VESD (Cai et al. 2018)	44.7	49.3	49.6	50.3	47.8	48.5	48.7	44.1	49.2	48.8	48.1
LIM-S (Xiong et al. 2019)	55.9	42.9	61.2	54.0	60.3	47.5	43.2	66.3	69.1	62.6	56.3
Trailer (Wang et al. 2020)	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
MINI-Net† (Hong et al. 2020)	80.6	68.3	78.2	81.8	78.1	65.8	75.8	75.0	80.2	65.5	73.2
SL-Module (Xu et al. 2021)	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
SA (Badamdorj et al. 2021)	83.4	64.7	84.4	86.5	70.3	67.5	66.9	68.1	95.0	60.8	74.8
SA <sup>+</sup> † (Badamdorj et al. 2021)	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	97.4	67.5	76.3
Joint-VA† (Badamdorj et al. 2021)	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	97.4	67.5	76.3
TCG† (Ye et al. 2021)	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
PLD-VHD (Wei et al. 2022)	84.5	80.9	70.3	72.5	76.4	87.2	71.9	74.0	74.4	79.1	77.1
UniVTG (Lin et al. 2023)	83.9	85.1	89.0	80.1	84.6	87.0	70.9	91.7	73.5	69.3	81.0
MH-DETR (Xu et al. 2024)	86.1	79.4	84.3	85.8	81.2	83.9	74.3	82.7	86.5	71.6	81.6
MIM† (Li et al. 2023)	84.4	85.8	91.3	73.9	83.1	87.1	80.1	78.2	80.3	79.6	82.4
UMT† (Liu et al. 2022b)	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
VCSJT† (Zhou et al. 2024)	87.5	80.7	88.6	76.6	83.6	91.0	77.6	93.3	88.9	80.0	84.8
QD-DETR (Moon et al. 2023b)	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
<b>LD-DETR (我们的模型)</b>	83.1	90.3	91.5	82.5	87.7	83.5	79.0	90.4	86.1	77.3	85.1
UVCOM (Xiao et al. 2024)	87.6	91.6	91.4	86.7	86.9	86.9	76.9	92.3	87.4	75.6	86.3
QD-DETR† (Moon et al. 2023b)	87.6	91.7	90.2	88.3	84.1	88.3	78.7	91.2	87.8	77.7	86.6
CG-DETR (Moon et al. 2023a)	86.9	88.8	94.8	87.7	86.7	89.6	74.8	93.3	89.2	75.9	86.8
TR-DETR† (Sun et al. 2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
TaskWeave (Yang et al. 2024)	88.2	90.8	93.3	87.5	87.0	82.0	80.9	92.9	89.5	81.2	87.3
LLMEPET (Jiang et al. 2024)	90.8	92.0	93.8	81.5	87.5	86.0	79.6	96.2	88.0	79.0	87.4
TR-DETR (Sun et al. 2024)	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
CDIM (黎金宇 2024)	85.5	95.8	90.3	90.0	88.4	88.1	79.2	97.1	88.0	80.5	88.3

表 3: TV-Sum 上的精彩片段检测结果。† 表示使用音频模态进行训练。

设置	视频时刻检索					精彩片段检测	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR (Lei, Berg, and Bansal 2021)	53.63 $\pm$ 1.59	35.78 $\pm$ 1.25	54.99 $\pm$ 1.33	31.28 $\pm$ 1.07	32.09 $\pm$ 1.01	36.41 $\pm$ 0.40	56.92 $\pm$ 0.98
M-DETR + 提取对齐	<b>55.56<math>\pm</math>1.01</b>	<b>36.72<math>\pm</math>1.43</b>	<b>56.42<math>\pm</math>0.76</b>	<b>31.43<math>\pm</math>0.99</b>	<b>32.66<math>\pm</math>0.73</b>	<b>37.16<math>\pm</math>0.36</b>	<b>58.62<math>\pm</math>1.19</b>
BM-DETR (Jung et al. 2023)	60.90 $\pm$ 0.79	44.10 $\pm$ 1.43	60.91 $\pm$ 0.96	39.03 $\pm$ 1.46	<b>38.93<math>\pm</math>0.98</b>	-	-
BM-DETR + 提取对齐	<b>61.04<math>\pm</math>1.02</b>	<b>44.23<math>\pm</math>1.04</b>	<b>61.53<math>\pm</math>0.98</b>	<b>39.12<math>\pm</math>1.45</b>	38.68 $\pm$ 1.09	-	-
QD-DETR (Moon et al. 2023b)	61.98 $\pm$ 0.55	47.30 $\pm$ 0.69	62.03 $\pm$ 0.43	41.96 $\pm$ 0.66	41.42 $\pm$ 0.28	38.92 $\pm$ 0.30	62.05 $\pm$ 0.86
QD-DETR + 提取对齐	<b>64.10<math>\pm</math>0.42</b>	<b>48.53<math>\pm</math>0.70</b>	<b>63.71<math>\pm</math>0.37</b>	<b>43.55<math>\pm</math>0.59</b>	<b>42.80<math>\pm</math>0.46</b>	<b>39.58<math>\pm</math>0.40</b>	<b>63.21<math>\pm</math>1.05</b>
CG-DETR (Moon et al. 2023a)	65.92 $\pm$ 0.22	50.44 $\pm$ 0.54	65.50 $\pm$ 0.22	45.62 $\pm$ 0.68	44.76 $\pm$ 0.26	40.34 $\pm$ 0.20	65.12 $\pm$ 0.64
CG-DETR + 提取对齐	<b>66.11<math>\pm</math>0.79</b>	<b>51.18<math>\pm</math>0.89</b>	<b>65.65<math>\pm</math>0.61</b>	<b>46.12<math>\pm</math>0.74</b>	<b>45.23<math>\pm</math>0.61</b>	<b>40.50<math>\pm</math>0.23</b>	<b>65.92<math>\pm</math>1.04</b>
TR-DETR (Sun et al. 2024)	66.56 $\pm$ 1.06	50.13 $\pm$ 0.89	65.70 $\pm$ 0.79	45.10 $\pm$ 0.78	44.33 $\pm$ 0.51	40.88 $\pm$ 0.19	65.54 $\pm$ 0.45
TR-DETR - LGMA	62.28 $\pm$ 1.08	46.99 $\pm$ 1.21	62.38 $\pm$ 0.72	62.16 $\pm$ 1.41	41.56 $\pm$ 1.12	39.16 $\pm$ 0.32	62.13 $\pm$ 0.46
TR-DETR + 提取对齐	66.31 $\pm$ 0.60	49.92 $\pm$ 0.73	65.33 $\pm$ 0.74	44.14 $\pm$ 0.81	43.60 $\pm$ 0.94	<b>40.89<math>\pm</math>0.26</b>	<b>65.79<math>\pm</math>1.02</b>
TR-DETR - LGMA + 提取对齐	<b>67.14<math>\pm</math>0.43</b>	<b>51.17<math>\pm</math>0.37</b>	<b>66.21<math>\pm</math>0.24</b>	<b>45.57<math>\pm</math>0.42</b>	<b>44.89<math>\pm</math>0.32</b>	40.77 $\pm$ 0.31	65.33 $\pm$ 0.71
UVCOM (Xiao et al. 2024)	64.36 $\pm$ 0.44	50.21 $\pm$ 1.06	63.99 $\pm$ 0.27	45.52 $\pm$ 0.68	44.77 $\pm$ 0.53	39.85 $\pm$ 0.21	63.82 $\pm$ 1.05
UVCOM + 提取对齐	<b>66.39<math>\pm</math>0.37</b>	<b>51.54<math>\pm</math>0.22</b>	<b>65.24<math>\pm</math>0.49</b>	<b>46.18<math>\pm</math>0.38</b>	<b>45.39<math>\pm</math>0.29</b>	<b>40.72<math>\pm</math>0.14</b>	<b>65.20<math>\pm</math>0.73</b>

表 4: 提取对齐作为一种即插即用的方法，可以帮助多个模型获得更好的结果。LGMA 是局部-全局多模态比对 (Local-Global Multi-Modal Alignment) 的缩写，是 TR-DETR (Sun et al. 2024) 模型中使用的一种方法。每列中每个基准模型上的最佳结果以**粗体**突出显示。

批次大小/队列长度	GPU 内存	视频时刻检索					精彩片段检测	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
更大的批次大小								
bs = 32, ql = 0	1,350	66.19 $\pm$ 1.41	49.74 $\pm$ 1.70	65.27 $\pm$ 1.11	44.12 $\pm$ 1.41	43.81 $\pm$ 1.19	40.72 $\pm$ 0.17	65.19 $\pm$ 1.23
bs = 64, ql = 0	2,140	66.89 $\pm$ 0.75	50.07 $\pm$ 0.82	65.73 $\pm$ 0.39	44.53 $\pm$ 0.54	43.98 $\pm$ 0.50	40.99 $\pm$ 0.29	65.86 $\pm$ 0.84
bs = 128, ql = 0	3,328	66.81 $\pm$ 0.70	50.93 $\pm$ 0.46	65.93 $\pm$ 0.46	45.00 $\pm$ 0.43	44.53 $\pm$ 0.33	40.92 $\pm$ 0.20	65.77 $\pm$ 1.26
bs = 256, ql = 0	5,772	66.62 $\pm$ 0.59	49.56 $\pm$ 0.82	65.27 $\pm$ 0.51	43.46 $\pm$ 0.66	43.22 $\pm$ 0.83	40.76 $\pm$ 0.16	65.38 $\pm$ 0.52
bs = 512, ql = 0	10,530	65.20 $\pm$ 1.92	47.38 $\pm$ 2.31	63.49 $\pm$ 1.85	41.35 $\pm$ 2.33	40.90 $\pm$ 2.00	40.58 $\pm$ 0.17	64.94 $\pm$ 1.16
bs = 1024, ql = 0	20,956	19.15 $\pm$ 13.66	8.70 $\pm$ 8.05	26.29 $\pm$ 10.96	9.45 $\pm$ 6.48	11.73 $\pm$ 6.34	27.13 $\pm$ 5.20	37.18 $\pm$ 10.86
bs = 2048, ql = 0	内存不足	-	-	-	-	-	-	-
提取对齐（我们的方法）								
bs = 32, ql = 0	1,312	66.53 $\pm$ 0.96	49.90 $\pm$ 0.31	65.70 $\pm$ 0.89	44.57 $\pm$ 0.48	44.07 $\pm$ 0.13	40.52 $\pm$ 0.24	64.73 $\pm$ 0.95
bs = 32, ql = 96	1,318	66.53 $\pm$ 0.32	49.34 $\pm$ 0.33	65.61 $\pm$ 0.85	44.13 $\pm$ 0.35	43.77 $\pm$ 0.49	40.50 $\pm$ 0.16	64.66 $\pm$ 0.24
bs = 32, ql = 480	1,236	67.11 $\pm$ 1.49	50.48 $\pm$ 1.03	<b>66.28<math>\pm</math>1.03</b>	45.10 $\pm$ 0.51	44.18 $\pm$ 0.41	40.92 $\pm$ 0.23	65.91 $\pm$ 0.97
bs = 32, ql = 2016	1,458	66.76 $\pm$ 0.57	50.48 $\pm$ 0.40	65.73 $\pm$ 0.47	44.83 $\pm$ 0.68	44.22 $\pm$ 0.57	40.81 $\pm$ 0.39	65.09 $\pm$ 0.70
bs = 32, ql = 8160	1,364	66.74 $\pm$ 0.84	<b>50.81<math>\pm</math>0.50</b>	66.07 $\pm$ 0.42	<b>45.30<math>\pm</math>0.52</b>	44.53 $\pm$ 0.34	40.76 $\pm$ 0.19	<b>65.95<math>\pm</math>0.42</b>
bs = 32, ql = 3,2736	1,620	67.00 $\pm$ 0.75	50.71 $\pm$ 1.10	66.10 $\pm$ 0.67	45.06 $\pm$ 0.99	44.51 $\pm$ 0.80	<b>41.00<math>\pm</math>0.21</b>	65.64 $\pm$ 0.53
bs = 32, ql = 13,1040	2,440	<b>67.14<math>\pm</math>0.59</b>	50.63 $\pm$ 0.48	65.91 $\pm$ 0.32	45.21 $\pm$ 0.60	<b>44.67<math>\pm</math>0.30</b>	<b>41.00<math>\pm</math>0.23</b>	65.86 $\pm$ 0.82

表 5: 与使用更大的批次大小相比，使用提取对齐可以节省 GPU 内存并获得更好的结果。下表显示了使用不同批次大小或队列长度、使用更大的批次大小和使用提炼对齐时的 GPU 内存使用情况和训练结果。“bs”表示批次大小，“ql”表示队列长度。实验在具有 2,4564MiB GPU 内存的 NVIDIA GeForce RTX 4090 上使用 TR-DETR (Sun et al. 2024) 作为基准模型进行。每列中的最佳结果以**粗体**突出显示。此表显示了与图 7 相同的一组实验。

提取系数	视频时刻检索					精彩片段检测	
	R1		mAP			>=Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
0.0	66.28 $\pm$ 1.38	49.91 $\pm$ 0.44	65.64 $\pm$ 0.74	44.60 $\pm$ 0.68	44.23 $\pm$ 0.27	40.81 $\pm$ 0.21	65.03 $\pm$ 0.78
0.1	67.20 $\pm$ 0.64	50.50 $\pm$ 0.78	66.20 $\pm$ 0.40	44.83 $\pm$ 0.64	44.50 $\pm$ 0.60	40.85 $\pm$ 0.20	65.58 $\pm$ 0.89
0.2	67.15 $\pm$ 0.99	50.50 $\pm$ 0.26	66.28 $\pm$ 0.61	45.24 $\pm$ 0.39	44.63 $\pm$ 0.40	40.98 $\pm$ 0.18	66.00 $\pm$ 0.65
0.3	67.07 $\pm$ 0.59	50.62 $\pm$ 0.42	66.19 $\pm$ 0.54	45.29 $\pm$ 0.52	44.61 $\pm$ 0.49	<b>41.14<math>\pm</math>0.16</b>	65.78 $\pm$ 1.26
0.4	<b>67.59<math>\pm</math>0.71</b>	<b>50.89<math>\pm</math>0.82</b>	<b>66.48<math>\pm</math>0.37</b>	<b>45.50<math>\pm</math>0.58</b>	<b>44.86<math>\pm</math>0.50</b>	41.09 $\pm$ 0.22	<b>66.32<math>\pm</math>0.47</b>
0.5	66.76 $\pm$ 0.89	50.46 $\pm$ 0.62	66.16 $\pm$ 0.54	45.06 $\pm$ 0.50	44.66 $\pm$ 0.28	40.83 $\pm$ 0.37	65.56 $\pm$ 0.62
0.6	66.61 $\pm$ 0.43	50.55 $\pm$ 0.59	65.89 $\pm$ 0.54	45.32 $\pm$ 0.56	44.68 $\pm$ 0.49	40.65 $\pm$ 0.20	64.55 $\pm$ 0.67
0.7	66.59 $\pm$ 0.28	50.52 $\pm$ 0.81	65.71 $\pm$ 0.22	45.14 $\pm$ 0.49	44.63 $\pm$ 0.32	41.05 $\pm$ 0.19	65.43 $\pm$ 0.42
0.8	65.72 $\pm$ 0.94	49.55 $\pm$ 1.09	64.87 $\pm$ 0.57	44.05 $\pm$ 1.51	43.62 $\pm$ 1.09	40.60 $\pm$ 0.13	64.80 $\pm$ 0.50
0.9	65.25 $\pm$ 0.70	49.47 $\pm$ 0.33	64.75 $\pm$ 0.22	44.17 $\pm$ 0.36	43.69 $\pm$ 0.33	40.23 $\pm$ 0.34	63.88 $\pm$ 1.25
1.0	62.83 $\pm$ 0.82	47.26 $\pm$ 0.85	63.08 $\pm$ 0.49	42.54 $\pm$ 0.82	42.09 $\pm$ 0.58	39.37 $\pm$ 0.21	61.87 $\pm$ 0.39

表 6: 该表显示了提取对齐中提取系数对多模态比齐的影响。每列中每个基线上的最佳结果以**粗体**突出显示。此表显示了与图 8 相同的一组实验。

对齐的影响。其中，当提取系数 = 0.0 时，相当于不启用提取。可以看出，当使用一定程度的提取时，不同训练样本之间的重叠语义信息被多模态对齐考虑在内，模型的效果变得更好。

**卷积融合器的消融实验** 表 7 展示了卷积融合器中卷积层数量的影响。视频中的运动信息往往蕴含在若干个局部片段中，卷积层可以有效的捕捉视频中的局部信

息。与 UVCOM (Xiao et al. 2024) (a-f) 相比，我们的方法 (g-m) 可以更好地利用卷积层提取局部特征的能力。在 UVCOM 中，随着卷积层数量的增加，模型的性能并没有变好。但在我们的方法中，随着卷积层数量的增加，模型的性能变得更好。与不使用卷积层 (g) 相比，仅添加一个卷积层 (h) 可以大大提升模型效果。随着加入的卷积层数量的增加，模型能够更有效地捕捉视频中的局部信息，效果逐渐提升直至达到峰值 (l)。

设置	# Conv	视频时刻检索					精彩片段检测	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
UVCOM (Xiao et al. 2024)								
(a)	0	63.68 $\pm$ 0.43	49.29 $\pm$ 0.86	63.50 $\pm$ 0.46	43.81 $\pm$ 0.44	43.40 $\pm$ 0.41	39.40 $\pm$ 0.13	63.22 $\pm$ 0.46
(b)	1	64.36 $\pm$ 0.44	50.21 $\pm$ 1.06	63.99 $\pm$ 0.27	45.52 $\pm$ 0.68	44.72 $\pm$ 0.53	39.85 $\pm$ 0.21	63.82 $\pm$ 1.05
(c)	2	61.68 $\pm$ 0.59	47.51 $\pm$ 0.54	61.72 $\pm$ 0.74	42.66 $\pm$ 0.53	42.40 $\pm$ 0.47	38.64 $\pm$ 0.43	60.90 $\pm$ 0.81
(d)	4	61.89 $\pm$ 0.63	47.54 $\pm$ 0.85	61.63 $\pm$ 0.82	42.46 $\pm$ 0.78	42.46 $\pm$ 0.81	38.52 $\pm$ 0.13	60.54 $\pm$ 0.75
(e)	8	60.87 $\pm$ 0.23	46.95 $\pm$ 0.63	61.41 $\pm$ 0.32	42.33 $\pm$ 0.50	42.28 $\pm$ 0.23	38.39 $\pm$ 0.23	60.22 $\pm$ 0.92
(f)	16	61.46 $\pm$ 0.60	47.13 $\pm$ 0.42	61.76 $\pm$ 0.25	43.07 $\pm$ 0.62	42.57 $\pm$ 0.33	38.19 $\pm$ 0.23	59.80 $\pm$ 0.77
LD-DETR (我们的模型)								
(g)	0	67.07 $\pm$ 1.54	50.96 $\pm$ 2.17	65.79 $\pm$ 1.72	45.17 $\pm$ 2.10	44.18 $\pm$ 1.91	41.42 $\pm$ 0.09	66.30 $\pm$ 0.29
(h)	1	68.14 $\pm$ 0.26	52.01 $\pm$ 1.28	66.70 $\pm$ 0.38	46.24 $\pm$ 1.18	45.71 $\pm$ 0.96	41.70 $\pm$ 0.20	<b>67.41<math>\pm</math>0.95</b>
(i)	2	68.10 $\pm$ 0.63	52.04 $\pm$ 0.45	67.49 $\pm$ 0.78	47.01 $\pm$ 0.25	46.73 $\pm$ 0.32	41.65 $\pm$ 0.24	66.62 $\pm$ 0.62
(j)	4	68.77 $\pm$ 0.40	52.88 $\pm$ 0.66	68.00 $\pm$ 0.38	47.55 $\pm$ 0.66	47.21 $\pm$ 0.46	<b>41.83<math>\pm</math>0.15</b>	67.30 $\pm$ 1.33
(k)	8	68.52 $\pm$ 0.90	52.22 $\pm$ 1.06	67.70 $\pm$ 0.62	47.14 $\pm$ 0.90	47.28 $\pm$ 0.66	41.55 $\pm$ 0.24	66.41 $\pm$ 0.61
(l)	10	<b>69.01<math>\pm</math>1.09</b>	<b>53.19<math>\pm</math>0.38</b>	<b>68.43<math>\pm</math>0.83</b>	<b>48.25<math>\pm</math>0.59</b>	<b>47.93<math>\pm</math>0.39</b>	41.66 $\pm$ 0.15	66.80 $\pm$ 0.96
(m)	16	68.63 $\pm$ 0.97	52.89 $\pm$ 0.62	68.10 $\pm$ 0.64	47.56 $\pm$ 0.36	47.24 $\pm$ 0.46	41.79 $\pm$ 0.15	67.33 $\pm$ 0.42

表 7: 该表显示了卷积层数量对模型性能的影响。“# Conv”表示卷积层的数量。每列中每个基线上的最佳结果以**粗体**突出显示。可以注意到，在 LD-DETR 上，借助卷积融合器，仅添加一个卷积层就可以显著提高模型性能。随着卷积层数量的增加，模型的性能会越来越好，直到达到极限。然而，在 UVCOM (Xiao et al. 2024) 上，添加卷积层并不能帮助模型表现得更好。

设置	位置	视频时刻检索					精彩片段检测	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)	V2TE <sub>x</sub> 之前	58.72 $\pm$ 2.48	42.43 $\pm$ 3.08	58.45 $\pm$ 2.87	37.44 $\pm$ 2.88	37.14 $\pm$ 2.84	38.20 $\pm$ 0.62	60.39 $\pm$ 1.60
(b)	V2TE <sub>x</sub> & T2VE <sub>n</sub> 之间	66.64 $\pm$ 1.21	49.61 $\pm$ 1.43	65.84 $\pm$ 1.00	43.99 $\pm$ 1.19	43.51 $\pm$ 1.15	41.20 $\pm$ 0.19	66.36 $\pm$ 0.48
(c)	T2VE <sub>n</sub> & TrEn1 之间	68.87 $\pm$ 0.68	52.96 $\pm$ 0.32	68.07 $\pm$ 0.56	47.54 $\pm$ 0.41	47.51 $\pm$ 0.39	<b>41.81<math>\pm</math>0.13</b>	66.79 $\pm$ 0.50
(d)	TrEn1 & TrEn2 之间	<b>69.01<math>\pm</math>1.09</b>	<b>53.19<math>\pm</math>0.38</b>	<b>68.43<math>\pm</math>0.83</b>	<b>48.25<math>\pm</math>0.59</b>	<b>47.93<math>\pm</math>0.39</b>	41.66 $\pm$ 0.15	<b>66.80<math>\pm</math>0.96</b>
(e)	TrEn2 之后	67.62 $\pm$ 0.96	52.23 $\pm$ 0.53	67.17 $\pm$ 0.17	46.85 $\pm$ 0.23	46.57 $\pm$ 0.19	41.56 $\pm$ 0.28	66.26 $\pm$ 0.52

表 8: 该表显示了卷积块在卷积融合器中放置在不同位置时模型的性能。可以注意到，当卷积块放置在 Transformer 编码器 1 和 Transformer 编码器 2 之间时，模型性能最佳。在该表中，“V2TE<sub>x</sub>”表示 V2T 提取器，“T2VE<sub>n</sub>”表示 T2V 编码器，“TrEn1”表示 Transformer 编码器 1，“TrEn2”表示 Transformer 编码器 2。每列中的最佳结果以**粗体**突出显示。

表 8 和表 9 展示了卷积融合器上的消融实验。无论改变方法的顺序还是删除其中任何一个，模型的性能都会下降。

**循环解码器的消融实验** 图 1 展示了循环解码器使视频时刻检索更加准确。它可视化了每次循环时循环解码器输出对应的视频时刻检索结果。随着循环次数的增加，结果越来越接近真实值。图 9、表 10 和表 11 展示了循环解码器在多个模型上的表现。它展示了循环解码器作为一种即插即用的方法，可以提高多个模型的性能。通过循环解码器方法，多模态信息被更充分地解码。使用循环解码器后，模型的性能得到了很大的提高。

然而，当使用相同尺寸的更大的解码器时，模型很快就会过拟合。我们注意到，循环解码器在 UVCOM (Xiao et al. 2024) 上的效果不如在其他模型上明显。我们推测这是因为它的双分支模态内聚合 (Dual Branches Intra-Modality Aggregation) 影响了我们方法的性能。当我们删除片段文本对齐 (Clip-Text Alignment, CTA) 方法时，随着循环次数的增加，性能会变得更好。但是当我们删除插槽注意力 (Slot Attention, SA) 方法时，模型的性能会变得更好。



设置	V2TEx	T2VEn	TrEn1	ConBl	TrEn2	视频时刻检索					精彩片段检测	
						R1		mAP			>=Very Good	
						@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)		✓	✓	✓	✓	63.82 $\pm$ 0.66	49.24 $\pm$ 0.77	63.94 $\pm$ 0.11	44.65 $\pm$ 0.28	44.28 $\pm$ 0.26	39.95 $\pm$ 0.27	62.92 $\pm$ 1.02
(b)	✓		✓	✓	✓	68.17 $\pm$ 0.88	52.41 $\pm$ 0.96	67.11 $\pm$ 0.73	46.74 $\pm$ 0.96	46.53 $\pm$ 0.50	41.50 $\pm$ 0.25	66.27 $\pm$ 0.95
(c)	✓	✓		✓	✓	68.09 $\pm$ 0.65	52.61 $\pm$ 0.84	67.53 $\pm$ 0.49	47.37 $\pm$ 0.42	47.03 $\pm$ 0.17	41.43 $\pm$ 0.15	66.62 $\pm$ 0.43
(d)	✓	✓	✓		✓	67.07 $\pm$ 1.54	50.96 $\pm$ 2.17	65.79 $\pm$ 1.72	45.17 $\pm$ 2.10	44.18 $\pm$ 1.91	41.42 $\pm$ 0.09	66.30 $\pm$ 0.29
(e)	✓	✓	✓	✓		68.36 $\pm$ 0.31	52.10 $\pm$ 0.36	67.67 $\pm$ 0.17	47.33 $\pm$ 0.65	47.07 $\pm$ 0.36	41.30 $\pm$ 0.24	66.75 $\pm$ 0.92
(f)	✓	✓	✓	✓	✓	<b>69.01<math>\pm</math>1.09</b>	<b>53.19<math>\pm</math>0.38</b>	<b>68.43<math>\pm</math>0.83</b>	<b>48.25<math>\pm</math>0.59</b>	<b>47.93<math>\pm</math>0.39</b>	<b>41.66<math>\pm</math>0.15</b>	<b>66.80<math>\pm</math>0.96</b>

表 9: 该表显示了卷积融合器中每个部分的必要性。当所有部分都存在时, 模型表现最佳。如果删除其中任何一部分, 模型表现都会变差。在该表中, “V2TEx” 表示 V2T 提取器, “T2VEn” 表示 T2V 编码器, “TrEn1” 表示 Transformer 编码器 1, “ConBl” 表示卷积块, “TrEn2” 表示 Transformer 编码器 2。每列中的最佳结果以**粗体**突出显示。

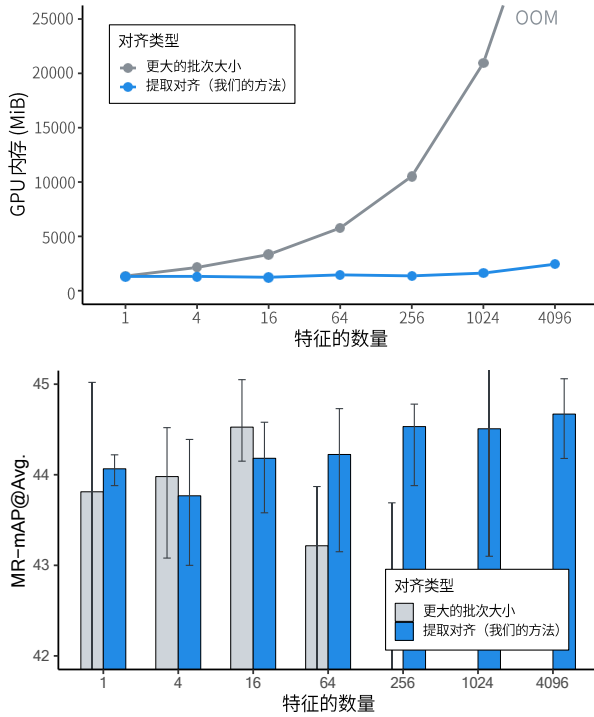


图 7: 提取对齐方法可以在不占用过多 GPU 内存的情况下引入更多样本进行对比学习, 随着队列长度的增加, 模型效果会越来越好。我们可视化了两种方法在对比学习所涉及的特征数量增加时的 GPU 内存占用情况以及在 QVHighlights 数据集上的结果。实验在 NVIDIA GeForce RTX 4090 上进行, 配备 2,4564MiB GPU 内存, 以 TR-DETR (Sun et al. 2024) 基准模型。x 轴表示对比学习所涉及的特征相对数量, 其中 x 轴 = 1 表示批次大小为 32, 队列长度为 0。“OOM” 表示内存不足。该图展示的实验与表 5 相同。

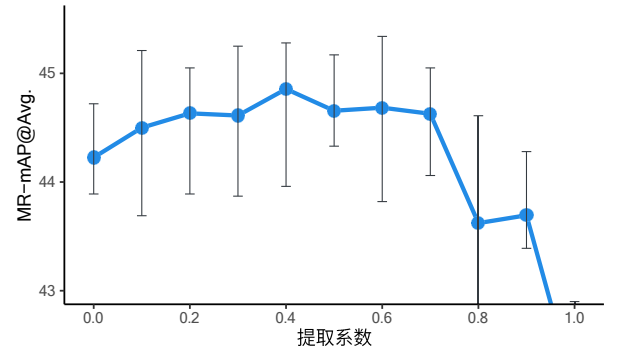


图 8: 该图显示了提取对齐中提取系数对多模态比齐的影响。该图显示了与表 6 相同的一组实验。

## 5 结论

在本文中, 我们提出了一种用于视频时刻检索和高光检测的模型 LD-DETR。我们首先提出了一种即插即用的方法提取对齐 (Distill Align), 它可以减轻重叠语义信息的影响。然后, 我们引入了卷积融合器 (Convolutional Fuser), 它更能够捕获多模态特征中的局部信息。最后, 我们提出了一种即插即用的方法循环解码器 (Loop Decoder), 它可以更充分地解码多模态信息而不会导致过度拟合。我们的方法的优越性和有效性已在四个公共数据集上得到证明。

## 参考文献

- Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In Proceedings of the IEEE international conference on computer vision, 5803–5812.
- Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint visual and audio learning for video highlight

解码器类型	解码器大小	视频时刻检索					精彩片段检测	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Moment-DETR (Lei, Berg, and Bansal 2021)								
更大的解码器	1	53.63 $\pm$ 1.59	35.78 $\pm$ 0.25	54.99 $\pm$ 1.33	31.28 $\pm$ 1.07	32.09 $\pm$ 1.01	<b>36.41<math>\pm</math>0.40</b>	<b>56.92<math>\pm</math>0.98</b>
	2	54.71 $\pm$ 2.35	35.84 $\pm$ 2.97	55.80 $\pm$ 1.82	31.39 $\pm$ 2.35	32.50 $\pm$ 1.99	36.17 $\pm$ 0.63	56.63 $\pm$ 1.03
	3	53.65 $\pm$ 0.73	35.15 $\pm$ 1.28	55.11 $\pm$ 0.68	30.34 $\pm$ 0.63	31.75 $\pm$ 0.73	36.04 $\pm$ 0.37	56.53 $\pm$ 0.98
	4	52.72 $\pm$ 0.97	35.14 $\pm$ 0.86	54.65 $\pm$ 0.59	30.64 $\pm$ 0.59	31.59 $\pm$ 0.50	35.78 $\pm$ 0.29	55.48 $\pm$ 0.63
循环解码器	1	53.63 $\pm$ 1.59	35.78 $\pm$ 0.25	54.99 $\pm$ 1.33	31.28 $\pm$ 1.07	32.09 $\pm$ 1.01	<b>36.41<math>\pm</math>0.40</b>	<b>56.92<math>\pm</math>0.98</b>
	2	55.69 $\pm$ 0.46	37.63 $\pm$ 0.55	<b>56.92<math>\pm</math>0.45</b>	32.93 $\pm$ 0.36	<b>33.86<math>\pm</math>0.16</b>	36.37 $\pm$ 0.41	56.88 $\pm$ 1.23
	3	54.93 $\pm$ 1.14	37.88 $\pm$ 1.28	56.19 $\pm$ 0.76	33.09 $\pm$ 0.81	33.66 $\pm$ 0.45	35.83 $\pm$ 0.16	55.30 $\pm$ 0.60
	4	<b>55.70<math>\pm</math>0.90</b>	<b>38.24<math>\pm</math>0.85</b>	56.74 $\pm$ 0.93	<b>33.19<math>\pm</math>0.70</b>	33.85 $\pm$ 0.66	36.10 $\pm$ 0.43	56.43 $\pm$ 0.54
BM-DETR (Jung et al. 2023)								
更大的解码器	1	60.90 $\pm$ 0.79	44.10 $\pm$ 1.43	60.91 $\pm$ 0.96	39.03 $\pm$ 1.46	38.93 $\pm$ 0.98	-	-
	2	61.24 $\pm$ 1.33	44.13 $\pm$ 1.42	61.25 $\pm$ 0.88	39.21 $\pm$ 1.06	38.96 $\pm$ 1.03	-	-
	3	60.72 $\pm$ 0.54	44.14 $\pm$ 1.00	61.14 $\pm$ 0.43	39.74 $\pm$ 1.05	39.18 $\pm$ 0.68	-	-
	4	60.82 $\pm$ 0.89	43.73 $\pm$ 0.78	61.20 $\pm$ 0.84	38.97 $\pm$ 0.80	38.83 $\pm$ 0.81	-	-
循环解码器	1	60.90 $\pm$ 0.79	44.10 $\pm$ 1.43	60.91 $\pm$ 0.96	39.03 $\pm$ 1.46	38.93 $\pm$ 0.98	-	-
	2	61.01 $\pm$ 0.70	<b>44.80<math>\pm</math>0.96</b>	61.68 $\pm$ 0.96	<b>40.34<math>\pm</math>0.93</b>	39.77 $\pm$ 0.75	-	-
	3	61.42 $\pm$ 1.20	44.70 $\pm$ 0.95	61.71 $\pm$ 1.06	39.75 $\pm$ 0.27	39.24 $\pm$ 0.63	-	-
	4	<b>61.62<math>\pm</math>0.73</b>	44.76 $\pm$ 1.48	<b>61.74<math>\pm</math>0.47</b>	40.14 $\pm$ 0.92	<b>39.81<math>\pm</math>0.87</b>	-	-
QD-DETR (Moon et al. 2023b)								
更大的解码器	1	61.98 $\pm$ 0.55	47.30 $\pm$ 0.69	62.03 $\pm$ 0.43	41.96 $\pm$ 0.66	41.42 $\pm$ 0.28	38.92 $\pm$ 0.30	62.05 $\pm$ 0.86
	2	62.59 $\pm$ 0.70	47.48 $\pm$ 0.79	62.02 $\pm$ 0.75	41.90 $\pm$ 0.63	41.37 $\pm$ 0.54	39.17 $\pm$ 0.21	62.18 $\pm$ 0.67
	3	61.73 $\pm$ 1.01	45.38 $\pm$ 1.63	60.45 $\pm$ 1.22	39.99 $\pm$ 1.75	39.57 $\pm$ 1.20	<b>39.24<math>\pm</math>0.23</b>	<b>62.99<math>\pm</math>0.74</b>
	4	60.96 $\pm$ 0.43	44.94 $\pm$ 1.03	59.92 $\pm$ 0.40	39.53 $\pm$ 0.76	39.04 $\pm$ 0.77	38.82 $\pm$ 0.27	61.39 $\pm$ 0.93
循环解码器	1	61.98 $\pm$ 0.55	47.30 $\pm$ 0.69	62.03 $\pm$ 0.43	41.96 $\pm$ 0.66	41.42 $\pm$ 0.28	38.92 $\pm$ 0.30	62.05 $\pm$ 0.86
	2	<b>62.80<math>\pm</math>0.80</b>	47.82 $\pm$ 0.53	62.97 $\pm$ 0.49	42.93 $\pm$ 0.61	42.26 $\pm$ 0.37	39.23 $\pm$ 0.13	62.83 $\pm$ 0.85
	3	62.49 $\pm$ 0.37	47.57 $\pm$ 0.44	<b>63.12<math>\pm</math>0.47</b>	42.68 $\pm$ 0.36	42.32 $\pm$ 0.32	38.94 $\pm$ 0.19	61.88 $\pm$ 0.43
	4	62.45 $\pm$ 0.86	<b>47.95<math>\pm</math>0.42</b>	63.03 $\pm$ 0.41	<b>43.20<math>\pm</math>0.08</b>	<b>42.48<math>\pm</math>0.24</b>	39.14 $\pm$ 0.15	62.67 $\pm$ 0.54
CG-DETR (Moon et al. 2023a)								
更大的解码器	1	65.92 $\pm$ 0.22	50.44 $\pm$ 0.54	65.50 $\pm$ 0.22	45.62 $\pm$ 0.68	44.76 $\pm$ 0.26	40.34 $\pm$ 0.20	65.12 $\pm$ 0.64
	2	65.81 $\pm$ 1.23	50.09 $\pm$ 0.54	64.40 $\pm$ 0.31	44.17 $\pm$ 0.51	43.71 $\pm$ 0.37	40.33 $\pm$ 0.25	65.10 $\pm$ 0.40
	3	65.30 $\pm$ 0.87	49.22 $\pm$ 0.95	63.50 $\pm$ 0.85	43.66 $\pm$ 0.45	42.85 $\pm$ 0.45	<b>40.45<math>\pm</math>0.20</b>	<b>65.61<math>\pm</math>0.50</b>
	4	63.87 $\pm$ 2.51	46.92 $\pm$ 4.35	61.63 $\pm$ 3.07	41.30 $\pm$ 3.54	40.78 $\pm$ 3.12	40.37 $\pm$ 0.11	65.03 $\pm$ 0.57
循环解码器	1	65.92 $\pm$ 0.22	50.44 $\pm$ 0.54	65.50 $\pm$ 0.22	45.62 $\pm$ 0.68	44.76 $\pm$ 0.26	40.34 $\pm$ 0.20	65.12 $\pm$ 0.64
	2	65.56 $\pm$ 0.49	50.58 $\pm$ 0.59	65.24 $\pm$ 0.44	45.29 $\pm$ 0.59	44.61 $\pm$ 0.39	40.10 $\pm$ 0.28	64.95 $\pm$ 0.79
	3	<b>66.26<math>\pm</math>0.64</b>	<b>51.72<math>\pm</math>0.53</b>	<b>65.74<math>\pm</math>0.39</b>	45.66 $\pm$ 0.27	45.05 $\pm$ 0.32	40.39 $\pm$ 0.24	65.53 $\pm$ 0.47
	4	66.22 $\pm$ 0.50	51.04 $\pm$ 0.36	66.01 $\pm$ 0.51	<b>45.95<math>\pm</math>0.52</b>	<b>45.23<math>\pm</math>0.35</b>	40.33 $\pm$ 0.13	65.50 $\pm$ 0.91
TR-DETR (Sun et al. 2024)								
更大的解码器	1	66.56 $\pm$ 1.06	50.13 $\pm$ 0.89	65.70 $\pm$ 0.79	45.10 $\pm$ 0.78	44.33 $\pm$ 0.51	40.88 $\pm$ 0.19	65.54 $\pm$ 0.45
	2	66.13 $\pm$ 1.09	49.54 $\pm$ 1.04	64.02 $\pm$ 1.63	43.86 $\pm$ 1.11	43.35 $\pm$ 0.91	40.61 $\pm$ 0.37	65.10 $\pm$ 1.00
	3	64.18 $\pm$ 1.22	46.79 $\pm$ 1.25	62.46 $\pm$ 0.96	42.12 $\pm$ 0.91	41.33 $\pm$ 0.85	<b>41.04<math>\pm</math>0.16</b>	65.28 $\pm$ 0.57
	4	64.71 $\pm$ 0.94	46.92 $\pm$ 1.35	62.69 $\pm$ 0.71	42.26 $\pm$ 1.25	41.54 $\pm$ 0.86	40.85 $\pm$ 0.10	<b>65.65<math>\pm</math>0.88</b>
循环解码器	1	66.56 $\pm$ 1.06	50.13 $\pm$ 0.89	65.70 $\pm$ 0.79	45.10 $\pm$ 0.78	44.33 $\pm$ 0.51	40.88 $\pm$ 0.19	65.54 $\pm$ 0.45
	2	66.59 $\pm$ 0.63	51.10 $\pm$ 0.40	65.97 $\pm$ 0.20	45.42 $\pm$ 0.20	44.90 $\pm$ 0.11	40.93 $\pm$ 0.29	65.60 $\pm$ 0.81
	3	<b>67.12<math>\pm</math>0.36</b>	51.08 $\pm$ 0.47	<b>66.35<math>\pm</math>0.46</b>	<b>45.58<math>\pm</math>0.47</b>	<b>44.92<math>\pm</math>0.33</b>	40.78 $\pm$ 0.44	65.37 $\pm$ 0.80
	4	66.94 $\pm$ 0.72	<b>51.40<math>\pm</math>0.51</b>	66.04 $\pm$ 0.69	45.22 $\pm$ 0.64	44.90 $\pm$ 0.34	40.85 $\pm$ 0.11	65.63 $\pm$ 0.69

表 10: 此页空间太小，写不开了。请参阅表格 11 了解详情。

解码器类型	解码器大小	视频时刻检索					精彩片段检测	
		R1		mAP			>=Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
UVCOM (Xiao et al. 2024)								
更大的解码器	1	64.49 $\pm$ 1.10	49.94 $\pm$ 0.72	64.03 $\pm$ 1.01	44.88 $\pm$ 0.53	44.39 $\pm$ 0.61	39.94 $\pm$ 0.18	64.03 $\pm$ 0.25
	2	64.37 $\pm$ 0.33	48.73 $\pm$ 0.73	62.85 $\pm$ 0.88	43.52 $\pm$ 1.16	42.73 $\pm$ 0.81	39.97 $\pm$ 0.16	63.42 $\pm$ 1.09
	3	63.21 $\pm$ 1.36	47.52 $\pm$ 1.81	61.61 $\pm$ 1.19	43.23 $\pm$ 1.35	42.36 $\pm$ 1.24	39.81 $\pm$ 0.13	63.44 $\pm$ 0.15
	4	62.68 $\pm$ 2.90	46.17 $\pm$ 3.85	60.74 $\pm$ 2.57	41.40 $\pm$ 2.81	40.99 $\pm$ 2.84	39.90 $\pm$ 0.48	63.72 $\pm$ 0.64
循环解码器	1	64.49 $\pm$ 1.10	49.94 $\pm$ 0.72	64.03 $\pm$ 1.01	44.88 $\pm$ 0.53	44.39 $\pm$ 0.61	39.94 $\pm$ 0.18	64.03 $\pm$ 0.25
	2	63.81 $\pm$ 0.37	49.50 $\pm$ 0.60	63.60 $\pm$ 0.38	44.44 $\pm$ 0.37	44.03 $\pm$ 0.24	39.81 $\pm$ 0.10	63.37 $\pm$ 0.38
	3	64.55 $\pm$ 0.69	50.04 $\pm$ 0.95	64.11 $\pm$ 0.54	45.41 $\pm$ 1.90	44.26 $\pm$ 0.56	39.94 $\pm$ 0.17	63.94 $\pm$ 0.46
	4	<b>64.65<math>\pm</math>0.61</b>	<b>50.33<math>\pm</math>0.55</b>	<b>64.17<math>\pm</math>0.46</b>	44.54 $\pm$ 0.44	44.18 $\pm$ 0.29	39.76 $\pm$ 0.18	63.69 $\pm$ 0.80
循环解码器 w/o SA	1	64.36 $\pm$ 0.44	50.21 $\pm$ 1.26	63.99 $\pm$ 0.27	<b>45.52<math>\pm</math>0.68</b>	<b>44.72<math>\pm</math>0.53</b>	39.85 $\pm$ 0.21	63.82 $\pm$ 1.05
	2	64.36 $\pm$ 0.54	50.06 $\pm$ 0.80	63.85 $\pm$ 0.30	45.15 $\pm$ 0.68	44.43 $\pm$ 0.51	39.88 $\pm$ 0.24	<b>64.19<math>\pm</math>0.59</b>
	3	64.00 $\pm$ 0.64	49.30 $\pm$ 1.22	63.73 $\pm$ 0.54	44.33 $\pm$ 1.00	43.93 $\pm$ 0.63	39.79 $\pm$ 0.20	63.54 $\pm$ 0.74
	4	63.39 $\pm$ 0.61	49.02 $\pm$ 0.50	63.32 $\pm$ 0.31	43.88 $\pm$ 0.52	43.20 $\pm$ 0.35	39.35 $\pm$ 0.18	62.59 $\pm$ 0.09
循环解码器 w/o CTA	1	63.96 $\pm$ 0.37	49.73 $\pm$ 0.42	63.65 $\pm$ 0.40	44.55 $\pm$ 0.47	44.12 $\pm$ 0.24	39.82 $\pm$ 0.20	63.63 $\pm$ 0.94
	2	<b>64.65<math>\pm</math>0.53</b>	50.09 $\pm$ 0.59	<b>64.17<math>\pm</math>0.72</b>	44.79 $\pm$ 0.87	44.42 $\pm$ 0.61	39.83 $\pm$ 0.17	63.90 $\pm$ 0.41
	3	64.62 $\pm$ 0.77	50.10 $\pm$ 0.40	64.12 $\pm$ 0.64	44.74 $\pm$ 0.44	44.25 $\pm$ 0.35	<b>39.98<math>\pm</math>0.15</b>	63.74 $\pm$ 0.36
	4	63.85 $\pm$ 0.78	49.55 $\pm$ 0.45	63.60 $\pm$ 1.13	44.87 $\pm$ 0.67	44.15 $\pm$ 0.82	35.72 $\pm$ 7.91	63.54 $\pm$ 0.67

表 11: 循环解码器作为一种即插即用的方法，可以帮助多个模型获得更好的结果，而不会像更大的解码器那样出现过拟合的风险。该表按基准模型对模型进行分类。对于循环解码器，解码器大小表示循环解码器循环的数量，对于更大的解码器，解码器大小表示解码器层数的倍数。插槽注意力 (Slot Attention, SA) 和片段文本对齐 (Clip-Text Alignment, CTA) 是 UVCOM (Xiao et al. 2024) 模型中使用的两种方法。每列中每类特征中的最佳结果以**粗体**突出显示。该表显示了与图 9 相同的一组实验。

detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 8127–8137.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Cai, S.; Zuo, W.; Davis, L. S.; and Zhang, L. 2018. Weakly-supervised video summarization using variational encoder-decoder and web prior. In Proceedings of the European conference on computer vision (ECCV), 184–200.

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video mamba suite: State space model as a versatile alternative for video understanding. arXiv preprint arXiv:2403.09626.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In Proceedings of the 2018 conference on empirical methods in natural language processing, 162–171.

Chen, S.; and Jiang, Y.-G. 2019. Semantic proposal for activity localization in videos via sentence query.

In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 8199–8206.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. 2019a. Finding moments in video collections using natural language. arXiv preprint arXiv:1907.12763.

Escorcia, V.; Soldan, M.; Sivic, J.; Ghanem, B.; and Russell, B. 2019b. Temporal localization of moments in video collections with natural language.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In Proceedings of the IEEE international conference on computer vision, 5267–5275.

Gao, J.; Sun, X.; Xu, M.; Zhou, X.; and Ghanem, B. 2021a. Relation-aware video reading comprehen-

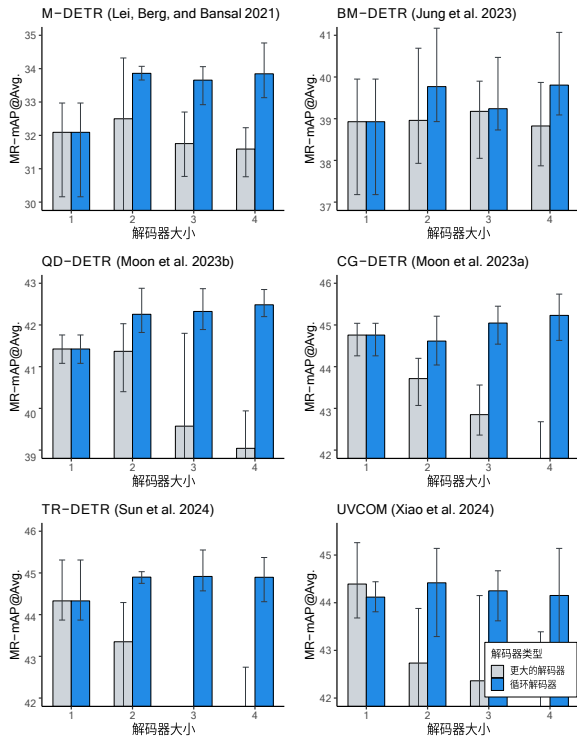


图 9: 循环解码器作为一种即插即用的方法, 可以帮助多个模型实现更好的结果, 而不会出现过拟合的风险。对于循环解码器, x 轴表示循环解码器循环的数量。对于更大的解码器解码器, x 轴表示解码器层数的倍数。此图显示了与表 10 和表 11 相同的一组实验。

sion for temporal language grounding. arXiv preprint arXiv:2110.05717.

Gao, P.; Zheng, M.; Wang, X.; Dai, J.; and Li, H. 2021b. Fast convergence of detr with spatially modulated co-attention. In Proceedings of the IEEE/CVF international conference on computer vision, 3621–3630.

Gygli, M.; Song, Y.; and Cao, L. 2016. Video2gif: Automatic generation of animated gifs from video. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1001–1009.

Han, D.; Cheng, X.; Guo, N.; Ye, X.; Rainer, B.; and Priller, P. 2023. Momentum cross-modal contrastive learning for video moment retrieval. IEEE Transactions on Circuits and Systems for Video Technology.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings

of the IEEE conference on computer vision and pattern recognition, 770–778.

Hoffmann, D. T.; Behrmann, N.; Gall, J.; Brox, T.; and Noroozi, M. 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 897–905.

Hong, F.-T.; Huang, X.; Li, W.-H.; and Zheng, W.-S. 2020. Mini-net: Multiple instance ranking network for video highlight detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, 345–360. Springer.

Huang, C.; Wu, Y.-L.; Shuai, H.-H.; and Huang, C.-C. 2024. Semantic Fusion Augmentation and Semantic Boundary Detection: A Novel Approach to Multi-Target Video Moment Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 6783–6792.

Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing where to focus: Event-aware transformer for video grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 13846–13856.

Ji, W.; Liang, R.; Zheng, Z.; Zhang, W.; Zhang, S.; Li, J.; Li, M.; and Chua, T.-s. 2023. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 23013–23022.

Jiang, Y.; Zhang, W.; Zhang, X.; Wei, X.; Chen, C. W.; and Li, Q. 2024. Prior Knowledge Integration via LLM Encoding and Pseudo Event Regulation for Video Moment Retrieval. arXiv preprint arXiv:2407.15051.

Jing, W.; Sun, A.; Zhang, H.; and Li, X. 2023. MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1387–1400. Toronto, Canada: Association for Computational Linguistics.

Jung, M.; Jang, Y.; Choi, S.; Kim, J.; Kim, J.-H.; and Zhang, B.-T. 2023. Overcoming Weak Visual-Textual Alignment for Video Moment Retrieval. arXiv preprint arXiv:2306.02728.



- Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.
- Lee, P.; and Byun, H. 2023. BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos. *arXiv preprint arXiv:2312.00083*.
- Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.
- Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* 16, 447–463. Springer.
- Li, J.; Zhang, F.; Lin, S.; Zhou, F.; and Wang, R. 2023. Mim: Lightweight multi-modal interaction model for joint video moment retrieval and highlight detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME), 1961–1966*. IEEE.
- Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1902–1910.
- Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2024. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36.
- Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.
- Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11235–11244.
- Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, 843–851.
- Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022a. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.
- Liu, W.; Mei, T.; Zhang, Y.; Che, C.; and Luo, J. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3707–3715.
- Liu, W.; Miao, B.; Cao, J.; Zhu, X.; Liu, B.; Nasim, M.; and Mian, A. 2024a. Context-Enhanced Video Moment Retrieval with Large Language Models. *arXiv preprint arXiv:2405.12540*.
- Liu, Y.; He, J.; Li, W.; Kim, J.; Wei, D.; Pfister, H.; and Chen, C. W. 2024b.  $R^2$ -Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. *arXiv preprint arXiv:2404.00801*.
- Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3042–3051.
- Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024c. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3855–3863.
- Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Ma, K.; Fang, H.; Zang, X.; Ban, C.; Zhou, L.; He, Z.; Li, Y.; Sun, H.; Feng, Z.; and Hou, X. 2024. Disentangle and denoise: Tackling context misalignment for video moment retrieval. *arXiv preprint arXiv:2408.07600*.
- Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; Wei, Y.; He, Z.; Li, Y.; and Sun, H. 2023. LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2798–2803.
- Ma, Y.; Yang, T.; Shan, Y.; and Li, X. 2022. Simvtp:

- Simple video text pre-training with masked autoencoders. arXiv preprint arXiv:2212.03490.
- Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF international conference on computer vision, 3651–3660.
- Moon, W.; Hyun, S.; Lee, S.; and Heo, J.-P. 2023a. Correlation-guided Query-Dependency Calibration in Video Representation Learning for Temporal Grounding. arXiv preprint arXiv:2311.08835.
- Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023b. Query-dependent video representation for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 23023–23033.
- Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional video grounding with dual contrastive learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2765–2775.
- Panta, L.; Shrestha, P.; Sapkota, B.; Bhattarai, A.; Manandhar, S.; and Sah, A. K. 2024. Cross-modal Contrastive Learning with Asymmetric Co-attention Network for Video Moment Retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 607–614.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, 8748–8763. PMLR.
- Regneri, M.; Rohrbach, M.; Wetzell, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics, 1: 25–36.
- Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 658–666.
- Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *nature*, 323(6088): 533–536.
- Soldan, M.; Xu, M.; Qu, S.; Tegner, J.; and Ghanem, B. 2021. Vlg-net: Video-language graph matching network for video grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 3224–3234.
- Song, Y.; Redi, M.; Vallmitjana, J.; and Jaimes, A. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In Proceedings of the 25th ACM international on conference on information and knowledge management, 659–668.
- Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5179–5187.
- Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. TR-DETR: Task-Reciprocal Transformer for Joint Moment Retrieval and Highlight Detection. arXiv preprint arXiv:2401.02309.
- Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, 776–794. Springer.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser,  [\$\square\$](#) ; and Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.
- Wang, L.; Liu, D.; Puri, R.; and Metaxas, D. N. 2020. Learning trailer moments in full-length movies with co-contrastive attention. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, 300–316. Springer.
- Wang, R.; Feng, J.; Zhang, F.; Luo, X.; and Luo, Y. 2024. Modality-aware Heterogeneous Graph for Joint Video Moment Retrieval and Highlight Detection. IEEE Transactions on Circuits and Systems for Video Technology.
- Wang, T.; Zhang, J.; Zheng, F.; Jiang, W.; Cheng, R.; and Luo, P. 2023. Learning grounded vision-language representation for versatile understanding in untrimmed videos. arXiv preprint arXiv:2303.06378.
- Wang, W.; Huang, Y.; and Wang, L. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 334–343.
- Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022a. Anchor detr: Query design for transformer-based detector. In Proceedings of the AAAI conference on artificial intelligence, volume 36, 2567–2575.

- Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.
- Wei, F.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Learning pixel-level distinctions for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3073–3082.
- Xiao, S.; Chen, L.; Zhang, S.; Ji, W.; Shao, J.; Ye, L.; and Xiao, J. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2986–2994.
- Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18709–18719.
- Xiong, B.; Kalantidis, Y.; Ghadiyaram, D.; and Grauman, K. 2019. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1258–1267.
- Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7970–7979.
- Xu, Y.; Sun, Y.; Zhai, B.; Jia, Y.; and Du, S. 2024. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.
- Yan, S.; Xiong, X.; Nagrani, A.; Arnab, A.; Wang, Z.; Ge, W.; Ross, D.; and Schmid, C. 2023. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13623–13633.
- Yang, J.; Wei, P.; Li, H.; and Ren, Z. 2024. Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18308–18318.
- Yao, T.; Mei, T.; and Rui, Y. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 982–990.
- Yao, Z.; Ai, J.; Li, B.; and Zhang, C. 2021. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*.
- Ye, Q.; Shen, X.; Gao, Y.; Wang, Z.; Bi, Q.; Li, P.; and Yang, G. 2021. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7950–7959.
- Yu, S.; Cho, J.; Yadav, P.; and Bansal, M. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.
- Yu, Y.; Lee, S.; Na, J.; Kang, J.; and Kim, G. 2018. A deep ranking model for spatio-temporal highlight detection from a 360<sup>∘</sup> video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Yuan, L.; Tay, F. E. H.; Li, P.; and Feng, J. 2019a. Un-supervised video summarization with cycle-consistent adversarial LSTM networks. *IEEE Transactions on Multimedia*, 22(10): 2711–2722.
- Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019b. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32.
- Yuan, Y.; Mei, T.; and Zhu, W. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9159–9166.
- Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10287–10296.
- Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1247–1257.
- Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

- Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII* 14, 766–782. Springer.
- Zhang, M.; Yang, Y.; Chen, X.; Ji, Y.; Xu, X.; Li, J.; and Shen, H. T. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12669–12678.
- Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.
- Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; and Dong, H. 2020. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*.
- Zhou, S.; Zhang, F.; Wang, R.; Zhou, F.; and Su, Z. 2024. Subtask Prior-driven Optimized Mechanism on Joint Video Moment Retrieval and Highlight Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.
- 杨金福; 刘玉斌; 宋琳; and 闫雪. 2022. 基于显著特征增强的跨模态视频片段检索. *电子与信息学报*, 44(12): 4395–4404.
- 缪翌; 张卫锋; and 徐领. 2024. 基于 CLIP 的视频时刻检索预训练模型. *计算机应用研究*, 1–8.
- 蒋寻; 徐行; 沈复民; 王国庆; and 杨阳. 2023. 无模态融合的高效弱监督视频时刻检索算法. *北京航空航天大学学报*, 1–12.
- 陈卓; 杜昊; 吴雨菲; 徐童; and 陈恩红. 2020. 基于视觉–文本关系对齐的跨模态视频片段检索. *中国科学: 信息科学*, 50(06): 862–876.
- 黎金字. 2024. 基于跨模态信息交互的视频内容定位研究. 硕士论文, 中山大学.